PB83-191056

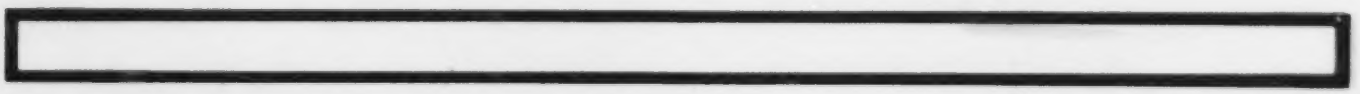
Variance Estimation Methods for the Emission Factor Equation

Technology Service Corp.
Santa Monica, CA

Prepared for

Environmental Protection Agency
Ann Arbor, MI

Jul 80



U.S. Department of Commerce
National Technical Information Service
NTIS

PB83-191056

Technology Service Corporation

2950 31st STREET • SANTA MONICA, CALIFORNIA 90405 • (213) 450-9755

TSC-PD-A231-2

VARIANCE ESTIMATION METHODS FOR THE EMISSION FACTOR EQUATION

Draft Final Report for

Contract No. 68-03-2836

Task 1

John Eldon
John Gins
Charles Stone

July 1980

Submitted to: Jay Wallace
U.S. Environmental Protection Agency
2565 Plymouth Road
Ann Arbor, Michigan 48105

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
		PB83 191056

4. TITLE AND SUBTITLE	5. REPORT DATE
Variance Estimation Methods for Emission Factor Equation.	July 1980
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
John Eldon, John Gins, Charles Stone	TSC-PD-A231-2
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Technology Service Corporation Santa Monica, California 90405	11. CONTRACT/GRANT NO. 68-03-2836
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
United States Environmental Protection Agency Office of Mobile Source Air Pollution Control Test and Evaluation Branch Ann Arbor, MI 48105	Technical 14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This report provides the steps involved in computing the United States Environmental Protection Agency's Mobile Source Emission Factor Equation (EFE) based on many variables. The EFE is used to estimate total emissions from motor vehicles representing the most recent model years.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Air Quality Algorithms Exhaust Emissions Mathematical Models Motor Vehicles Variables	Environmental Protection Agency Emission Factors Mobile Source Emission Model Variance Estimation Methods	

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
Release Unlimited	unclassified	71
	20. SECURITY CLASS *(This page)* unclassified	22. PRICE

EPA Form 2220-1 (9-73)

TABLE OF CONTENTS

Section

1. INTRODUCTION..... 1

1.1 OBJECTIVES OF THIS DOCUMENT..... 3
1.2 USE OF THIS DOCUMENT..... 3
1.3 SUMMARY OF THE METHOD..... 4
1.4 CONCLUSIONS AND RECOMMENDATIONS..... 7

2. PRELIMINARY COMPUTATIONS..... 10

2.1 COMPUTING THE MEAN AND COVARIANCE MATRIX..... 10
2.2 INTERPRETING THE INITIAL COMPUTATIONS..... 13
2.3 TEST FOR NORMALITY OF DATA..... 14
2.4 USING ESTIMATES OF THE MEAN AND VARIANCE..... 16

3. VARIANCE ESTIMATION METHODS..... 17

3.1 VARIANCE ESTIMATION FOR INDEPENDENT VARIABLES..... 17
3.2 GENERAL VARIANCE ESTIMATOR FOR CORRELATED NORMAL VARIABLES..... 18
3.2.1 Transformation of Variables..... 19
3.2.2 Computation of Variance..... 23
3.2.3 Use of the Variance Algorithm..... 25

4. EXAMPLE OF VARIANCE ESTIMATION..... 27

4.1 COMPUTATIONAL EXAMPLES..... 27
4.2 NUMERICAL EXAMPLES..... 31

REFERENCES..... 34

Appendices

A Statistical Tables..... 35

B Test for Bivariate Normality..... 46

1. INTRODUCTION

The U.S. Environmental Protection Agency's Mobile Source Emission Factor Equation (EFE) is based on many variables and is used to estimate composite (total) emissions from motor vehicles representing the twenty most recent model years. The EFE allows the air pollution modeler to consider explicitly the impacts of seven "component factors" in estimating total emissions of a fleet of vehicles. These factors include vehicle operating conditions (hot/cold operation, load, and speed), environmental factors (temperature and humidity), the vehicular age/usage mix, and the vehicle emission rates themselves.

The EFE was developed primarily through analyses of data collected during new vehicle certificationa and used vehicle surveillance emissions testing.

The overall equation for light duty vehicles is:

$$Enpstwx = SUM(Cipn*Min*Ripstwx*Aip*Ip*Uipw*Hip)$$

where: all lower case letters or subscripts

SUM( ) = summation over model year (i), from the calendar year for which emission factors are being calculated (i=n) to the calendar year 19 years previous (i=n-19).

Enpstwx = Composite emission factor in gm/mi for calendar year n, pollutant p, average speed s, ambient temperature t, fraction cold operation w, and fraction hot start operation x.

Cipn = The FTP (1975 Federal Test Procedure) mean emission factor for the ith model year light-duty vehicles during calendar year n, and for pollutant p.

Min = The fraction of annual travel by the ith model year LDVs during calendar year n.

Ripstwx = The temperature, speed, and hot/cold correction factor for the ith model year LDVs for pollutant p, average speed s, ambient temperature t, fraction cold operation w, and fraction hot start operation x.

Aip = The air-conditioning correction factor for the ith model year LDVs, for pollutant p.

Lp = The vehicle load correction factor for pollutant p.

Uipw = The trailer towing correction factor for the ith model year LDVs, for pollutant p, and for fraction of cold operation x.

Hip = The humidity correction factor for the ith model year LDVs, for pollutant p.

Any emission estimate made using the EFE will contain uncertainties, which arise from two major sources. First, there are uncertainties associated with the component factors which are contained in the EFE. Second, there undoubtedly exist physical factors which influence motor vehicle emissions, but which are not contained in the equation. Naturally, the emissions estimate can be only as good as the identified sources of variation within the data.

Currently, the emission factor estimation method provides only for the development of point estimates of composite emissions, based on point estimates of the component environmental and vehicle operation/emission factors. There is no provision for determining the overall uncertainty of the composite emission estimate or for computing a range of emission estimates, given the uncertainties (variances) of the component factors.

Under contract to the U.S. Environmental Protection Agency, Technology Service Corporation (TSC) has developed a set of procedures

for estimating the variance of composite motor vehicle emissions estimates from the known or estimated variances of the seven component environmental and vehicle operation factors. (The uncertainties associated with external physical factors (i.e., factors not in the EFE) are outside the scope of this study).

1.1 OBJECTIVES OF THIS DOCUMENT

The primary purpose of this document is to provide the EPA with a useful book which lists the steps involved in computing the EFE variance for a single model year. Each model of the twenty model year EFE's are assumed to be independent. Thus, the model variance over all twenty years will be the sum of the variances computed for each of the twenty model years. Specifically listed are the caveats associated with the use of the method described and the types of cases to which the method cannot be applied. This report also lists the types of assumptions which can be used to broaden the applicability of the procedure to additional situations.

1.2 USE OF THIS DOCUMENT

Because of the expected use of the information presented herein, this report has been organized in "guidebook" format. By following the methods described in this volume in the sequence in which they are covered, the reader should be able to estimate composite variance. The higher-dimensional cases will probably require the aid of a computer, but some of the simpler, low-dimensional cases will require only a hand calculator.

The specific organization of this document is as follows. Chapter 2 outlines the fundamental computations which must be completed before the variance can be estimated. The purpose of these calculations

is to develop a set of matrices and to compute a series of indices which will be needed in the actual variance computations. Chapter 3 describes two general methods for estimating variance. Section 3.1 discusses the simpler technique, which can be used only if the component factors are assumed (or shown) to be independent, while the rest of Chapter 3 addresses the more complex technique for mutually correlated, but normally distributed, variables. For problems involving four or more component factors, where the variance equation becomes unwieldy, we have presented a general algorithm for developing the appropriate multi-term equations. Finally, Chapter 4, which presents sample applications of the algorithm to the cases of two and three correlated normal variables, has been added to make the somewhat theoretical treatment of Chapter 3 more comprehensible and useful. The solutions to higher order problems, up to the full seven dimensions, follow the logic and structure of these simpler cases, although the number of terms is far greater. The most general case, which would involve correlated non-normal variables, is not treated in this document.

1.3 SUMMARY OF THE METHOD

Figure 1.1 is a flow chart illustrating the processes necessary to compute the variance in any given problem. The first step is to examine the problem in detail and to identify and count the relevant free parameters (component variables affecting the composite emission estimate in the Emission Factor Equation). For this list of variables, the mean vector and the covariance and correlation matrices are then

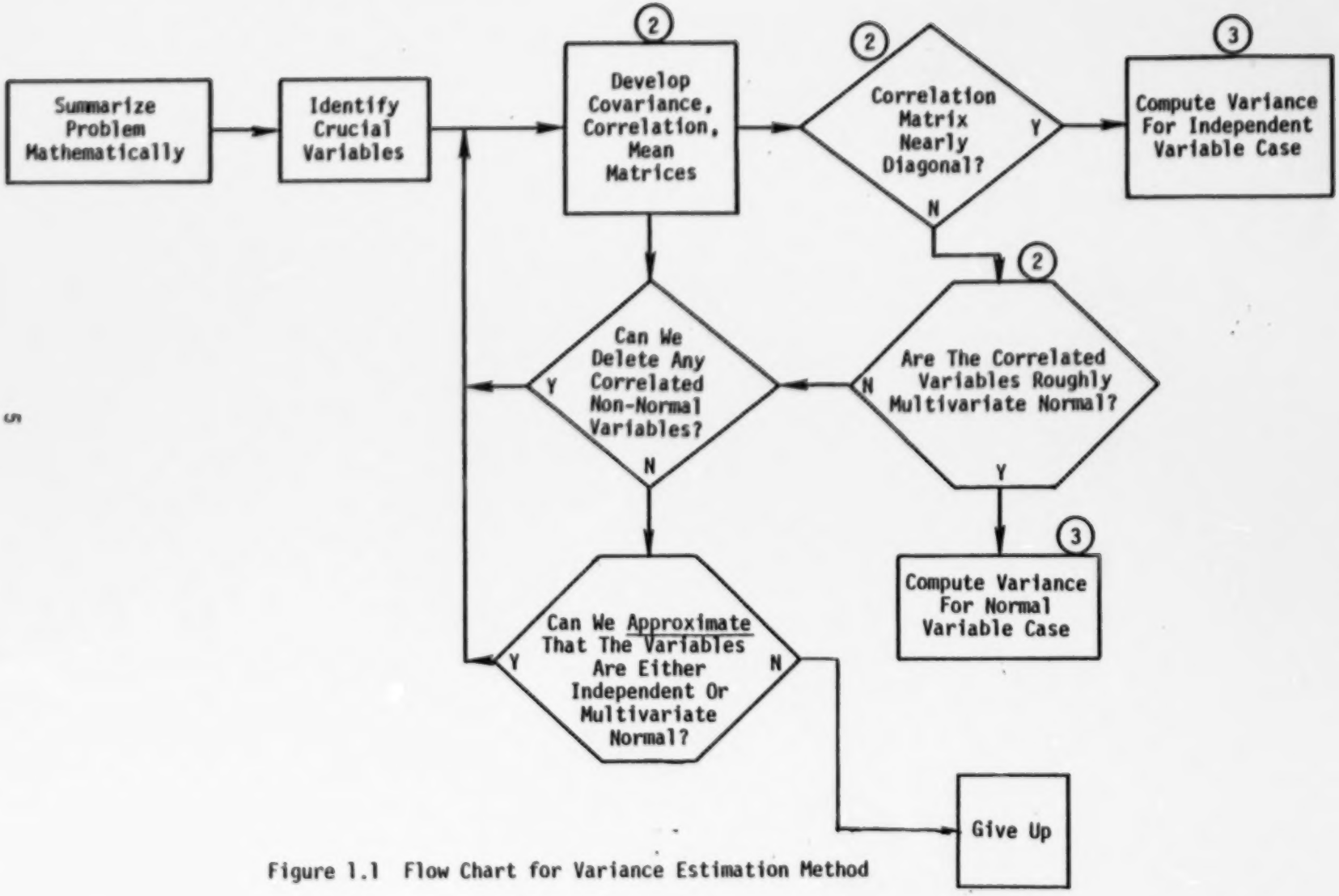


Figure 1.1 Flow Chart for Variance Estimation Method

developed, per the instructions in Chapter 2. From the correlation matrix, one can determine whether or not the component factors are mutually independent; the less correlated two factors are, the smaller the corresponding off-diagonal terms will be. A group of totally orthogonal (mutually independent) variables will yield a purely diagonal correlation matrix, i.e., one whose off-diagonal terms are all zero. The reader can test the matrix for independence of the variables, by comparing the size of the off-diagonal elements against calculated critical values.

If the off-diagonal elements are very small in absolute value, then the reader may assume independence of the component factors and proceed directly to compute the composite variance, as explained in Chapter 3. If two or more factors are very strongly correlated, it may be possible to combine them by regression into a single new factor, thereby eliminating the correlation. The result will then involve a reduced number of variables, all of which may be mutually independent.

If two or more component factors are significantly correlated, but not so highly that they can be combined, then the next step is to assess the normality of their distributions. The assessment procedure is sketched in Chapter 2, which also refers the reader to previous reports telling how to determine whether a given factor is normally distributed and whether two factors are bivariate normally distributed. If the condition of multivariate normality is met or can be assumed, then the user will use the large algorithm for computation of composite variance for the correlated n-dimensional, n-variate normal case. Note that this

procedure will also apply to the case in which some of the variables are correlated with one another, but independent of the others. Where two variables are nearly independent, their near-zero off-diagonal matrix terms will cause all related variance equation terms to be very small.

Where the variables are not multivariate normal, but are correlated, then one must take certain approximations in order to compute overall variance. Approximation methods are briefly discussed in Chapter 2.

1.4 CONCLUSIONS AND RECOMMENDATIONS

The statistical work reported in this document has demonstrated that one can compute the variance of a parameter from the variances of its component factors. However, the work has also shown that the problem is only tractable when the component factors are either independent or multi-variate normally distributed. When the factor variables are correlated but not normally distributed, there is no reasonably simple way to compute the composite variance.

With correlated component factors, the problem of computing overall variance can become "messy," involving numerous terms. Therefore, this report presents the actual variance equations only for the simplest cases (2 and 3 dimensions) and presents an algorithm for generating the higher-dimensional espressions, which can entail hundreds of terms. Dimension in this report is equivalent to factor (i.e., a 3-dimensional case means that only three factors are involved in the computation of an EFE for a single model year). The number of terms rises rapidly with the number of dimensions,

as shown in Table 1.1. Each of these terms are products of various coefficients and all these terms will need to be added together in order to derive the estimate of the mean or variance.

Correlations among the component factors also complicate the computation of the mean of the product of these factors, which is not simply the product of respective means, as it would be in the uncorrelated case.

The specific computations of the mean and variance will not directly lead to a computation of a confidence interval about the mean, because the underlying distribution of the product has not been determined. However, using Monte Carlo simulation techniques, the underlying distribution for a given covariance matrix and multivariate normality of the variables could be approximated. This would require generating about one million standard normal random variables, placing them appropriately in the variance equation, and computing the resulting emission factor. The distribution of these random emission factors could then be found and appropriate confidence intervals could be readily developed.

TABLE 1.1 COMPLEXITY OF EQUATIONS*

Number of Dimensions (Component Factors)	Number of Original Terms**	
	Mean	Variance
2	6	21
3	24	320
4	120	14,520
5	720	519,120
6	5,040	25,406,640
7	40,320	812,871,360

* This table applies to the correlated, normally distributed variable case, the most complex one considered in this document. The number of terms is far smaller if some or all of the variables are mutually independent.

** This is the maximum number of terms, which will be required if all variables are correlated with each other.

2. PRELIMINARY COMPUTATIONS

The first step in the variance estimator method is a series of "preliminary" computations, the results of which will be needed for the composite variance estimator. This chapter outlines the necessary preliminary computations and tells how to interpret their results. This chapter also tells how to evaluate the normality and independence of the factor variables.

2.1 COMPUTING THE MEAN AND COVARIANCE MATRIX

The simplest computation is that of the mean vector, $\vec{\mu}$, whose values are the means of the individual component factors. For this purpose, the arithmetic mean of each variable is computed and entered in the corresponding position in the mean vector, $\vec{\mu}$ (i.e., the first entry, μ_1, is the mean of variable #1, the second, μ_2, is that of variable #2, etc.). The formula for estimating μ_i the i^{th} factor's arithmetic mean, $\hat{\mu}_i$, is:

$$\hat{\mu}_i = \frac{1}{n}\sum_k x_{ik} \qquad (2\text{-}1)$$

where n is the number of data points, x_{ik}, for the i^{th} variable, and k is a dummy counter, 1,...,n.[1]

The next step is to estimate the variance σ^2 and standard deviation σ for each component factor. The variances are estimated according to the standard equation:

$$\hat{\sigma}_i^2 = \frac{1}{n}\left[\sum_k x_{ik}^2 - \frac{1}{n}\left(\sum_k x_{ik}\right)^2\right] \qquad (2\text{-}2)$$

where σ_i is the estimated standard deviation of the i^{th} factor, and n is the number of useable data points in its distribution.[1] Variance and standard deviation routines are available in the standard computer statistical packages, such as SPSS. The number of useable data points, "n," requires further explanation. In all subsequent matrices, one must ensure that the data points of all component factors represent a common data set. Therefore, in preparing the data base, one must delete all data points which are missing the values for one or more of the component factors. Alternatively, one can retain data points for which all missing factor values can be estimated or computed from the other data values. Thus, if the data base includes a particular value for one of the component factors, it must also contain the corresponding actual or estimated values for <u>all</u> of the other factors.

The covariance matrix, $\vec{\Sigma}$, lists the variance of each individual variable, plus the covariance of each possible pair of variables. Standard computer packages, such as SPSS and BMD-P, include easy-to-use covariance matrix routines. The covariance will be estimated from the original data. The general formula for the (i,j) element (i^{th} row and j^{th} column) of the covariance matrix is:

$$\hat{\sigma}_{ij} = \frac{1}{n}\left[\sum_k x_{ik}\, x_{jk} - \frac{\sum_k x_{ik} \sum_k x_{jk}}{n}\right] \qquad (2\text{-}3)$$

where x_i is the i^{th} variable and x_j is the j^{th}, and n is the number of data points. [1] The matrix is symmetrical, such that the (j,i) element equals the (i,j), as is readily apparent from the formula. The (i,i) element, the variance of the i^{th} variable, is entered in the i^{th} row and the i^{th} column.

The next matrix required is the correlation matrix, R, whose elements are related to those of the covariance matrix by the simple equation:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i \sigma_j}} \cdot [1] \qquad (2\text{-}4)$$

In this case, the (i,j) element of the matrix is the correlation coefficient between the variables i and j. The diagonal elements (i,i) are one in this case, since each variable is perfectly correlated with itself.

Again, one must ensure that the sample sizes of all variables in the covariance matrix are equal. If any single observation is missing, then all corresponding observations of the other variables for that case must be deleted (i.e., if x_{ij}, the j^{th} observation of the i^{th} variable, is missing, then throw away the j^{th} observations for all other variables. The set of variables for the j^{th} observation is called the j^{th} case). Thus, when data are gathered to compute the covariance matrix $\hat{\Sigma}$, an observation for each variable and case must be obtained. In certain circumstances, however, a particular variable may have an excessive

number of missing data values (e.g., due to an instrument malfunction). By deleting all corresponding observations for the other variables, one may reduce the data set to an unacceptably small sample size. (Generally, a data base of fewer than 30 points is unacceptably small, and a sample size of 50 or more is preferred.) In such a case, instead of deleting so many data points, it may be preferable to assume that the "sparse-data" variable is independent of the others. Under this assumption, the covariances between the "sparse data" variable and the others will all be zero.

2.2 INTERPRETING THE INITIAL COMPUTATIONS

Calculation methods for computing the variance estimator vary according to the mutual correlations among the component factors. Therefore, in order to select the appropriate calculation procedure, one must first determine whether any of the factors are correlated, by testing the statistical significance of each correlation in the correlation matrix. If none of the correlations prove to be significant, then the simple procedure outlined in Section 3.1 can be used. However, if any of the variables are correlated with one another, then the more complex algorithms of Section 3.2 must be applied.

The off-diagonal elements of the correlation matrix are the best single measure of the mutual independences of the various component factors. If the (i,j) element of this matrix is either zero or "not significantly different from zero," then the i^{th} and j^{th} variables can be assumed to

be linearly independent. If all off-diagonal terms are zero, than all variables are linearly independent, and the Section 3.1 algorithm can be used. In practice, however, it is almost certain that some or all of the off-diagonal terms will be nonzero, in which case the next step is to test the significance of each deviation from zero. The simplest reliable test of linear independence is to compare each matrix element with its corresponding critical value. The critical value, a function of the sample size (number of data points), n, and the desired confidence level (usually either 95% or 99%), is given by $2/\sqrt{n}$ (at the 95% level) or $2.6/\sqrt{n}$ (at the 99% level). For this particular application, we recommend using the 99% confidence level, which will lead to an assumption of linear independence unless we are 99% certain that two variables are correlated. If a matrix element exceeds its critical value, then one cannot assume that the two corresponding variables are independent of each other, and must proceed with the more complex form of the variance estimator method. In contrast, if the (i,j) element is smaller than the corresponding critical value, then one can assume that the factors i and j are linearly independent.

2.3 TEST FOR NORMALITY OF DATA

If the analysis of independence indicates that two or more variables are correlated, one must then test these variables for multivariate normality. Unfortunately, convenient, efficient statistical tests are available only for univariate and bivariate normality, rather than general multivariate normality. The recommended procedure is as follows:

(1) Identify, among the seven factors, any variables which are mutually correlated. Test each correlated variable for normality, using the Wilks-Shapiro (Appendix C) or the other tests for normality. If any "correlated" variable is not normally distributed, the EFE variance cannot be computed unless one of the following conditions is met:

1. The non-normal "correlated" variable can be ignored (treated as a constant) in the EFE.

2. The non-normal "correlated" factor variable can actually be assumed independent of the other variables. This option is particularly attractive if the variable in question is only weakly correlated with the others.

3. The variable is only slightly nonnormal, and could therefore be assumed normal.

4. The variable is so strongly correlated with one or more others that its values can be predicted from those of the others, via regression analysis. In this case, a single variable can represent one or more other variables as well as itself, thereby reducing the total number of variables.

If two or more variables are "correlated nonnormals," then one of these four conditions must apply to each. If not, the overall variance cannot be estimated using the procedures in this document.

(2) If all single variables are normal, then test each pair of correlated variables for bivariate normality. For example, if the variables #1, 2, and 3 are all correlated with one another, test the variable pairs (1,2), (2,3), and (1,3) for bivariate normality. The standard test for bivariate normality is presented in Appendix B.

(3) Assess the status of the normality and correlations of the variables, to decide which procedure to observe. If all of the variables

are independent, then the Section 3.1 algorithm is appropriate. Similarly, if two variables are correlated, but one of the two can be eliminated (by regression) or ignored, then Section 3.1 still applies. In contrast, if some of the final variables are correlated, but normally distributed, then one must use the Section 3.2 algorithm to develop a variance estimate. Low-dimensional cases can be handled on a calculator, using the formulae presented in Chapter 4, while higher dimensional cases will require the aid of a computer.

2.4 USING ESTIMATES OF THE MEAN AND VARIANCE

In the following two chapters the theoretical mean vector and covariance vector are treated as known values. In practice, however, the means, variances and covariances are estimated from a sample and used as if they are the true parameters.

3. VARIANCE ESTIMATION METHODS

This chapter presents two methods for computing the variance of the EFE from those of its component factors. The first (Section 3.1) can be used only if all component factors are mutually independent, or nearly so, while the second (Section 3.2) can be used with correlated factors as well. However, the second procedure can be used only if all mutually correlated variables are multivariate normally distributed, while the first can be used irrespective of the distributions of the factors.

3.1 VARIANCE ESTIMATION FOR INDEPENDENT VARIABLES

The general two-dimensional formula for estimating the overall variance from the variances of two independent component factor variables is:

$$\mathrm{VAR}(X_1X_2) = \left(\mu_1^2 + \sigma_1^2\right)\left(\mu_2^2 + \sigma_2^2\right) - (\mu_1\mu_2)^2$$

$$= \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2 \quad .$$

The mean value of the product X_1X_2 is simply the product of the mean, $\mu_1\mu_2$.

Similarly, the three-dimensional case mean is $\mu_1\mu_2\mu_3$, while the variance is:

$$\mathrm{VAR}(X_1X_2X_3) = \left(\mu_1^2 + \sigma_1^2\right)\left(\mu_2^2 + \sigma_2^2\right)\left(\mu_3^2 + \sigma_3^2\right) - (\mu_1\mu_2\mu_3)^2$$

$$= \mu_1^2\sigma_2^2\sigma_3^2 + \mu_2^2\sigma_1^2\sigma_3^2 + \mu_3^2\sigma_1^2\sigma_2^2 + \mu_1^2\mu_2^2\sigma_3^2 + \mu_1^2\mu_3^2\sigma_2^2 + \mu_2^2\mu_3^2\sigma_1^2$$

$$+ \sigma_1^2\sigma_2^2\sigma_3^2 \quad .$$

The seven-dimensional equation follows the same pattern, viz:

$$\mathrm{VAR}(X_1 \cdots X_7) = \left(\mu_1^2 + \sigma_1^2\right)\left(\mu_2^2 + \sigma_2^2\right)\cdots\left(\mu_7^2 + \sigma_7^2\right) - (\mu_1 \cdots \mu_7)^2 \quad .$$

These equations make no assumptions about the distributions of the component factors, but they do assume mutual independence.

3.2 GENERAL VARIANCE ESTIMATOR FOR CORRELATED NORMAL VARIABLES

This section outlines an algorithm that will enable the user to calculate the mean and variance of the product of up to seven correlated random component factor variables. The procedure entails the transformation of the original factor variables to a set of orthogonal (independent), normally distributed, normalized (mean = 0 and standard deviation = 1) variables in n-dimensional space, where "n" is both the original number of factor variables under consideration and the number of transformed variables. This orthogonalization and normalization greatly simplify the algebra in the later steps of the algorithm.

3.2.1 Transformation of Variables

The original factor variables are assumed to have mean vector $\vec{\mu}$ and covariance matrix $\vec{\Sigma}$. To perform the transformation, one must first compute the elements, σ_{ij}, of $\vec{\Sigma}$. This covariance matrix can be written as the cross-product of a lower-triangular matrix $\vec{L}$ and its transpose $\vec{L}^T$, which can be derived using the Cholesky decomposition routine given in the next paragraph. (A lower-triangular matrix is one whose off-diagonal terms above and to the right of the main diagonal are zero. See Table 3.1). The lower triangular format of the matrix $\vec{L}$ reflects the transformation's consideration of the seven factor variables in a randomly selected, but fixed sequence. Each variable is expressed in terms of a new orthogonal variable, plus those of the previously considered factor variables. Thus, the first transformation involves expressing X_1, the first factor variable, in terms of U_1, the first transformation variable, where U_1 is chosen to align with X_1. X_2 is then expressed in terms of U_1 and U_2 where the direction of U_2 is chosen such that the plane of U_1 and U_2 contains X_2. Similarly X_3's transformation involves terms in U_1, U_2, and U_3, etc.

The following iterative algorithm due to Stewart [2] is used to convert from these elements to the actual elements (ℓ_{ij}) of $\vec{L}$.

TABLE 3.1 FORMAT OF MATRIX $\vec{L}$

ℓ_{11}	0	0	0	0	0	0
ℓ_{21}	ℓ_{22}	0	0	0	0	0
ℓ_{31}	ℓ_{32}	ℓ_{33}	0	0	0	0
			. . .			
ℓ_{71}	ℓ_{72}	ℓ_{73}	ℓ_{74}	ℓ_{75}	ℓ_{76}	ℓ_{77}

(1) Initialize all ℓ_{ij} values to the corresponding value σ_{ij} thus

$$\ell_{ii} = \sigma_i^2$$

and

$$\ell_{ij} = \sigma_{ij}$$

(2) For i = 1 to n

(2.1) first for j = 1 to i - 1, do:

$$\ell_{ij} = \ell_{jj}^{-1}\left(\ell_{ij} - \sum_{k=1}^{j-1} \ell_{ik}\ell_{jk}\right) .$$

(2.2) then set

$$\ell_{ii} = \sqrt{\ell_{ii} - \sum_{k=1}^{i-1} \ell_{ik}^2} .$$

These equations have been written in computer programming format, with the "old" values to the right of the equals sign, and the "new" values to the left. Separate equations are presented for the off-diagonal and diagonal elements.

The seven original variables, X_1, X_2,...,X_7, can each be written as a sum of new independent normal variables U_1, U_2,...,U_7 as follows:

$$X_1 = \mu_1 + \ell_{11}U_1$$

$$X_2 = \mu_2 + \ell_{21}U_1 + \ell_{22}U_2$$

$$\vdots$$

$$X_7 = \mu_7 + \ell_{71}U_1 + \ell_{72}U_2 + \ell_{73}U_3 + \ell_{74}U_4 + \ell_{75}U_5 + \ell_{76}U_6 + \ell_{77}U_7$$

In general, $X_n = \mu_n + \sum_{i=1}^{n} \ell_{ni}U_i$, where all ℓ_{ij} are elements of $\hat{L}$ matrix. [3]

The variance of $X_1 \cdot X_2 \cdot X_3 \cdots X_7$ can be written as:

$$\text{Var}(X_1 \cdot X_2 \cdots X_7) = E\left(X_1^2 \cdot X_2^2 \cdots X_7^2\right) - (E(X_1 \cdot X_2 \cdots X_7))^2$$

$$= E((X_1 \cdot X_2 \cdots X_7)^2) - (E(X_1 \cdot X_2 \cdots X_7))^2$$

$$= E\left(A_7^2\right) - (E(A_7))^2$$

where $A_7 = (\mu_1 + \ell_{11}U_1)(\mu_2 + \ell_{21}U_1 + \ell_{22}U_2) \cdots (\mu_7 + \ell_{71}U_1 + \ell_{72}U_2 + \cdots + \ell_{77}U_7)$ and $E(Z)$ is the expectated value of Z. The values of $\mu_1, \mu_2, \ldots, \mu_7$ and $\ell_{11}, \ell_{21}, \ldots, \ell_{77}$ are constants dependent on the populations for $X_1, \ldots, X_7$ and $U_1, U_2, \ldots, U_7$ are random variables. Since $U_1, \ldots, U_7$ are all independent and normal with mean 0 and standard deviation 1, then A_7 and A_7^2 can be rewritten algebraically as the sums of constant terms and products of $E(\mu_i^n)$. From statistical theory [4] we know $E(U_i) = 0$ $E(U_i^3) = E(U_i^5) = E(U_i^7) = E(U_i^9) = E(U_i^{11}) = E(U_i^{13}) = 0$ and $E(U_i^2) = 1$, $E(U_i^4) = 3$, $E(U_i^6) = 15$, $E(U_i^8) = 105$, $E(U_i^{10}) = 945$, $E(U_i^{12}) = 10395$, and $E(U_i^{14}) = 135135$.

The key to the problem involves finding the various unique products of U_1 through U_7, evaluating their expectation values, and multiplying by the appropriate constant terms of $\mu_1, \ldots, \mu_7$ and $\ell_{11}, \ldots, \ell_{77}$.

With a product of 2 or 3 variables it is easy to multiply all the terms, take the expected values, and sum the results to compute the variance. However, the problem becomes increasingly more difficult as more variables are used. Table 1.1 lists the number of terms that would be added together for each A_n and A_n^2. A_n is used in computing the means, while A_n^2 is used in computing the variances.

3.2.2 Computation of Variance

The best way to simplify the variance computation is to add together coefficients of the various unique combinations of $U_1,\ldots,U_7$. This is done by determining whether each "new" combination of terms has occurred before. Since we must keep track of the exponents n_i of each of the U_i terms, we will define a set, S, such that for each element $S(n_1,n_2,\ldots,n_7)$,

$$S(n_1,n_2,\ldots,n_7) = 0 \text{ if } U_1^{n_1} U_2^{n_2} \cdots U_7^{n_7} \text{ has never occurred before}$$

$$S(n_1,n_2,\ldots,n_7) = m > 0 \text{ if } U_1^{n_1} U_2^{n_2} \cdots U_7^{n_7} \text{ has occurred before.}$$

The value m is the location indicator for the element a_m of the set A. For each element a_m of A

$$a_m = a_s(n_1,\ldots,n_7) = \sum L_j(n_1,\ldots,n_7)$$

where $L_j(n_1,\ldots,n_7)$ is the coefficient of the j^{th} occurrence of a term of $U_1^{n_1} \cdots U_7^{n_7}$.

Let us define an indicator function $I(n_1,\ldots,n_7)$ such that

$$I(n_1,\ldots,n_7) = \begin{cases} 0 \text{ if any } n_i \text{ is odd or if } n_i = 0 \text{ for all } i \\ 1 \text{ otherwise} \end{cases}$$

and a function $F(U_i^n)$ such that

$$F(U_i^n) = \begin{cases} 0 \text{ if } n \text{ is odd or } n < 0 \\ 1 \text{ if } n = 0 \\ F(U_i^{n-2}) \cdot (n-1) \text{ otherwise} \end{cases}$$

Thus $F(U_i^2) = 1$, $F(U_i^4) = 3$, $F(U_i^6) = 15$, $F(U_i^8) = 105$, $F(U_i^{10}) = 945$, $F(U_i^{12}) = 10395$, and $F(U_i^{14}) = 135135$. The expected value of $(X_1 \cdot X_2 \cdot X_3 \cdots X_7)$ is then:

$$E(X_1 \cdot X_2 \cdot X_3 \cdots X_7) = \sum_J F\left(U_1^{n_1}\right)F\left(U_2^{n_2}\right)\cdots F\left(U_7^{n_7}\right) \cdot a_{s(\vec{j})} + \mu_1 \mu_2 \mu_3 \cdots \mu_7$$

where J is the set of vectors $\vec{j} = (n_1,\ldots,n_7)$ such that $s(\vec{j}) \neq 0$ is an element of the set S and $I(\vec{j}) = 1$.

The variance of $(x_1 \cdot x_2 \cdots x_7)$ is

$$\mathrm{Var}(x_1 \ldots x_7) = \sum_{J'} \sum_K F\left(U_1^{n_1+m_1}\right)F\left(U_2^{n_2+m_2}\right)\ldots F\left(U_7^{n_7+m_7}\right) \cdot a_{s(\vec{j})} \cdot a_{s(\vec{k})} \cdot G(\vec{j},\vec{k}) - (E(x_1 \cdot x_2 \cdots x_7) - \mu_1 \cdot \mu_2 \cdots \mu_7)^2$$

where J' is the set of vectors $\vec{j} = (n_1,\ldots,n_7)$ such that $s(\vec{j}) \neq 0$ is an element of the set S and $n_i > 0$ for some i where $1 \leq i \leq 7$ and where K is the set of vectors $\vec{k} = (m_1,\ldots,m_7)$ such that $\vec{k}$ was a vector that already was used under the outermost summation sign. In other words if there was an intrinsic order to the $s(\vec{j}) \neq 0$, then $0 < s(\vec{k}) \leq s(\vec{j})$. The function $G(\vec{j},\vec{k})$ is defined as follows:

$$G(\vec{j},\vec{k}) \begin{cases} 1 \text{ is } \vec{j} = \vec{k} \\ \\ 2 \text{ otherwise} \end{cases}$$

Thus $G(\vec{j},\vec{k})$ indicates whether a particular term is a binary combination or a square term. By defining the double summation in this way we have cut the number of additions almost in half.

In Chapter 4 we will carry out the above algorithm for the product of two terms and the product of three terms. In the next section we will describe how the elements of the set S can be ordered to minimize the number of calculations required. We will also describe how this algorithm can be used for a combination of independent and dependent elements.

3.2.3 Use of the Variance Algorithm

Much work can be saved if the elements $s(n_1,\ldots,n_7)$ of the set S, which was described in the previous section, are ordered correctly. We propose the use of the following mapping function, which puts the elements of S into a useable order. Let

$$M(n_1,\ldots,n_7) = 8(7(6(5(4(3H(n_7) + H(n_6)) + H(n_5))$$

$$+ H(n_4)) + H(n_3)) + H(n_2)) + H(n_1) + 1$$

where H( ) is a function which places the odd elements after the even elements. H( ) is defined as follows:

$$H(n_j) = \text{Int}\left(\frac{8-j}{2}\right) - (-1^{n_j})\text{Int}\left(\frac{n_i+1}{2}\right)$$

where Int( ) means "take the integer part of the real number." Let $s(n_1,\ldots,n_7)$ be the $M(n_1,\ldots,n_7)^{th}$ element of the set S.

After set S is completely set up, it can be compressed to set S', such that if $S(n_1,\ldots,n_7) > 0$, then $S(n_1,\ldots,n_7)$ is an element of S'. By keeping track of the starting position in S' of the odd and even values of n_i, where i is the current variable of interest and where $n_j = 0$ for all $j > i$, one can skip whole possible combinations of powers of U_i simply because $n_i + n_j$ is odd.

This algorithm can be used for independent variables as well, because if x_i is independent of all other x_j, $j \neq i$, then $\sigma_{ij} = 0$ for all $j \neq i$ and the corresponding $\ell_{ij} = 0$. Thus the computation of the variance is automatically simplified in all cases involving independent variables.

4. EXAMPLE OF VARIANCE ESTIMATION

This chapter outlines the variance estimation procedure for the simplest cases, which are those involving two or three dimensions. The basic procedure is easily extended to higher-dimensional cases, but the number of terms soon becomes very large, making a computer necessary for the computation.

4.1 COMPUTATIONAL EXAMPLES

The simplest case involves two variables, which we will label X_1 and X_2. Each X_i is normally distributed, with mean μ_i and standard deviation σ_i. Likewise, each X_i has an expectation value which is simply μ_i. Similarly, standard statistical theory tells us that the expectation value of each X_i^2 is $\mu_i^2 + \sigma_i^2$.

Before computing expectation values, it is useful to express the variables X_i in terms of two new independent normal variables, U_1 and U_2. Both U_i have means of 0 and standard deviations of 1. The U_i are defined such that

$$X_1 = \mu_1 + \sigma_1 U_1 = \mu_1 + \ell_{11} U_1$$

$$X_2 = \mu_2 + \ell_{21} U_1 + \ell_{22} U_2 \quad .$$

The value of X_1X_2 is then $(\mu_1 + \ell_{11}U_1)(\mu_2 + \ell_{21}U_1 + \ell_{22}U_2)$, which can be written as: $C_1 + C_2\, U_1^2$ + [other terms], where C_1 and C_2 are constants and all "other terms" involve odd powers of U_1 and/or U_2 and therefore have expectation values of zero. Here, $C_1 = \mu_1\mu_2$ and $C_2 = \ell_{11}\ell_{21}$. The expectation value of X_1X_2 is then:

$$E(X_1X_2) = C_1 + C_2\, EU_1^2 = C_1 + C_2 = \mu_1\mu_2 + \ell_{11}\ell_{21} \quad .$$

To find the variance of X_1X_2 we must evaluate $X_1^2X_2^2 = (\mu_1 + \ell_{11}U_1)^2 \cdot (\mu_2 + \ell_{21}U_1 + \ell_{22}U_2)^2$. Again taking advantage of the fact that the expectation value of any odd power of any U_i is zero, we write the expansion of this product as:

$$X_1^2X_2^2 = C_1 + C_2U_1^2 + C_3U_2^2 + C_4U_1^4 + C_5U_1^2U_2^2 + [\text{other terms}] \quad .$$

Then $E(X_1^2X_2^2) = C_1 + C_2 + C_3 + 3C_4 + C_5$, since the expectation value of each $U_i^2 = 1$, and $E(U_i^4) = 3$.

Since

$$C_1 = \mu_1^2\mu_2^2$$

$$C_2 = \mu_1^2\ell_{21}^2 + \mu_2^2\ell_{11}^2 + 4\mu_1\mu_2\ell_{11}\ell_{21}$$

$$C_3 = \mu_1^2\ell_{22}^2$$

$$C_4 = \ell_{11}^2\ell_{21}^2$$

$$C_5 = \ell_{11}^2\ell_{22}^2$$

$$E\left(X_1^2X_2^2\right) = \mu_1^2\mu_2^2 + \mu_1^2\ell_{21}^2 + \mu_1^2\ell_{22}^2 + \mu_2^2\ell_{11}^2 + 4\mu_1\mu_2\ell_{11}\ell_{21}$$

$$+ 3\ell_{11}^2\ell_{21}^2 + \ell_{11}^2\ell_{22}^2 \quad .$$

For two dimensions, the covariance matrix is

$$L = \begin{pmatrix} \sigma_1 & 0 \\ \sigma_2\rho_{12} & \sigma_2\sqrt{1-\rho_{12}^2} \end{pmatrix}$$

where ρ_{12} is the correlation between X_1 and X_2. Therefore, $\ell_{12} = 0$, $\ell_{11} = \sigma_1$, and $\ell_{21} = \sigma_{21}/\sigma_1$, because $\rho_{12} = \rho_{21} = \sigma_{21}/\sigma_1\sigma_2$.

Substituting back into the above equations, we have:

$$E\left(X_1X_2\right) = \mu_1\mu_2 + \ell_{11}\ell_{21}$$

$$= \mu_1\mu_2 + \sigma_1 \cdot \frac{\sigma_{21}}{\sigma_1} = \mu_1\mu_2 + \sigma_{21} \quad . \qquad (4\text{-}1)$$

Then the variance is given by:

$$\mathrm{Var}\left(X_1X_2\right) = E\, X_1^2X_2^2 - (E\, X_1X_2)^2$$

$$= \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + 2\mu_1\mu_2\sigma_{21} + \sigma_1^2\sigma_2^2 + \sigma_{12}^2 \quad . \qquad (4\text{-}2)$$

If X_1 and X_2 are <u>independent</u>, then $\sigma_{21} = \sigma_{12} = 0$, and

$$\mathrm{Var}(X_1X_2) = \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2 \quad ,$$

which is clearly equivalent to the 2-variable independent case treated in Section 3.1

The next most complex case involves three correlated, normal variables. The number of terms is far higher than in the simpler 2-dimensional case. Although it is feasible to perform the 3-dimensional variance estimation using a calculator, it would be much more efficient to set up a simple computer program to compute the variance for this case. The principal advantage of the computer is that it can keep track of all the numerous terms.

The 3-dimensional variance is computed as follows:

$$\mathrm{Var}(X_1X_2X_3) = E((X_1X_2X_3)^2) - (E(X_1X_2X_3))^2$$

The 3-dimensional transformation is given by:

$$X_1 = \mu_1 + \ell_{11}U_1$$
$$X_2 = \mu_2 + \ell_{21}U_2 + \ell_{22}U_2$$
$$X_3 = \mu_3 + \ell_{31}U_1 + \ell_{32}U_2 + \ell_{33}U_3$$

where

$$\ell_{11} = \sigma_1$$

$$\ell_{21} = \sigma_{21}/\sigma_1$$

$$\ell_{22} = \sqrt{\sigma_2^2 - \sigma_{21}^2/\sigma_1^2}$$

$$\ell_{31} = \sigma_{31}/\sigma_1$$

$$\ell_{32} = (\sigma_{32} - \sigma_{31}\sigma_{21}/\sigma_1^2)/\sqrt{\sigma_2^2 - \sigma_{21}^2/\sigma_1^2}$$

$$\ell_{33} = \sqrt{\sigma_3^2 - \sigma_{31}^2/\sigma_1^2 - (\sigma_{32} - \sigma_{31}\sigma_{21}/\sigma_1^2)^2/(\sigma_2^2 - \sigma_{21}^2/\sigma_1^2)}$$

The expectation value and variance are then:

$$E(X_1X_2X_3) = \mu_1\mu_2\mu_3 + \mu_2\sigma_{31} + \sigma_{21}\mu_3 + \mu_1\sigma_{32} \qquad (4\text{-}3)$$

$$\begin{aligned} Var(X_1X_2X_3) = {} & (\mu_1^2+\sigma_1^2)(\mu_2^2\sigma_3^2+\mu_3^2\sigma_2^2+\sigma_2^2\sigma_3^2) \\ & +2\mu_1\mu_2(\sigma_{21}\sigma_3^2+\sigma_{31}^2\sigma_{32})+4\sigma_{21}\sigma_{32}\sigma_{31}^2 \\ & +2\mu_1\mu_3(\sigma_{21}\sigma_{32}+2\sigma_2^2\sigma_{31}^2)+\sigma_2^2\sigma_{31}^2 \\ & +\sigma_1^2(\mu_2\mu_3+\sigma_{32})^2+2\mu_2\mu_3\sigma_{21}\sigma_{31}^2 \\ & +7\sigma_{21}^2\sigma_{31}^2/\sigma_1^2+\sigma_{21}^2\sigma_3^2 \end{aligned} \qquad (4\text{-}4)$$

4.2 NUMERICAL EXAMPLES

For sake of example let X, be a normal random variate with mean 1.6 and variance 0.16, let X_2 be a noraml random variable with mean 0.9 and variance 0.09, and let X_3 be a normal random variable with mean 2.5 and variance 0.25. The results of calculating the variance and mean of the

product of X_1 and X_2 under various correlation conditions are given in Table 4.1. The mean is a function of the correlation coefficient as well as the variance.

The results of calculating the mean and product of the three variables are given in Table 4.2. Both the mean and variance depend on the correlation coefficients.

TABLE 4.1 MEAN AND VARIANCE ESTIMATORS FOR THE PRODUCT OF TWO VARIABLES

Variable	Correlation	Mean	Variance
X_1	----	1.6	0.16
X_2	----	0.9	0.09
X_1X_2	0.0	1.44	0.374
X_1X_2	0.25	1.47	0.462
X_1X_2	-0.25	1.41	0.289
X_1X_2	0.5	1.5	0.551
X_1X_2	-0.5	1.38	0.205
X_1X_2	0.75	1.53	0.642
X_1X_2	-0.75	1.35	0.123
X_1X_2	1.00	1.56	0.734
X_1X_2	-1.00	1.32	0.043

TABLE 4.2 MEAN AND VARIANCE ESTIMATORS FOR THE PRODUCT OF THREE VARIABLES

Variable	Correlation of X_1X_2	X_1X_3	X_2X_3	Mean	Variance
X_1	--	--	--	1.6	0.16
X_2	--	--	--	0.9	0.09
X_3	--	--	--	2.5	0.25
$X_1X_2X_3$	0	0	0	3.6	2.952
$X_1X_2X_3$	0.5	0.5	0.5	3.96	3.212
$X_1X_2X_3$	-0.5	-0.5	-0.5	3.24	2.870
$X_1X_2X_3$	0	-0.5	0	3.51	2.881
$X_1X_2X_3$	0	0	-0.5	3.48	2.899
$X_1X_2X_3$	0.5	0	0	3.75	2.996
$X_1X_2X_3$	0	0.5	0	3.69	3.025
$X_1X_2X_3$	0	0	0.5	3.72	3.007
$X_1X_2X_3$	0.5	0.5	0	3.84	3.098
$X_1X_2X_3$	0	0.5	0.5	3.81	3.101
$X_1X_2X_3$	0.5	0	0.5	3.87	3.087
$X_1X_2X_3$	-0.5	-0.5	0	3.87	3.087
$X_1X_2X_3$	0	-0.5	-0.5	3.39	2.849
$X_1X_2X_3$	-0.5	0	-0.5	3.33	2.893
$X_1X_2X_3$	0.5	-0.5	-0.5	3.54	2.834
$X_1X_2X_3$	0.5	0.5	-0.5	3.72	2.985
$X_1X_2X_3$	0.5	-0.5	0.5	3.78	2.967
$X_1X_2X_3$	0.5	-0.5	0	3.66	2.900
$X_1X_2X_3$	0.5	0	-0.5	3.63	2.907
$X_1X_2X_3$	-0.5	-0.5	0.5	3.48	2.866
$X_1X_2X_3$	-0.5	0.5	-0.5	3.42	2.920
$X_1X_2X_3$	-0.5	0.5	0	3.54	2.957
$X_1X_2X_3$	-0.5	0	0.5	3.57	2.929
$X_1X_2X_3$	0	-0.5	0.5	3.63	2.914
$X_1X_2X_3$	0	0.5	-0.5	3.57	2.950

REFERENCES

[1] Afifi, A, and Azen, S Statistical Analysis, A Computer Oriented Approach, Academic Press, New York 1972.

[2] Stewart, G. Introduction to Matrix Computations, Academic Press, New York 1973.

[3] Graybill, F. An Introduction to Linear Statistical Models, Volume 1, McGraw-Hill, New York 1961.

[4] Hoel, P., Port, S. and Stone, C. Introduction to Probability Theory, Houghton Mifflin, Boston 1971.

APPENDIX A

STATISTICAL TABLES

These tables are referred to from Appendix B.

Table A-1a Table for Testing Skewness

(One-tailed percentage points of the distribution of $\sqrt{b_1} = g_1 = m_3/m_2^{3/2}$)*

Size of Sample n	Percentage Points		Standard Deviation	Size of Sample n	Percentage Points		Standard Deviation
	5%	1%			5%	1%	
25	0.711	1.061	0.4354	100	0.389	0.567	0.2377
30	0.662	0.986	.4052	125	0.350	0.508	.2139
35	0.621	0.923	.3804	150	0.321	0.464	1961
40	0.587	0.870	.3596	175	0.298	0.430	.1820
45	0.558	0.825	.3418	200	0.280	0.403	.1706
50	0.534	0.787	.3264				
60	0.492	0.723	.3009	250	0.251	0.360	.1531
70	0.459	0.673	.2806	300	0.230	0.329	.1400
80	0.432	0.631	.2638	350	0.213	0.305	.1298
90	0.409	0.596	.2498	400	0.200	0.285	.1216
100	0.389	0.567	.2377	450	0.188	0.269	.1147
				500	0.179	0.255	.1089

* Since the distribution of $\sqrt{b_1}$ is symmetrical about zero, the percentage points represent 10% and 2% two-tailed values.

Table A-1b Table for Testing Kurtosis

(Percentage points of the distribution of $b_2 = m_4/m_2^2$)*

Size of Sample n	Percentage Points				Size of Sample n	Percentage Points			
	Upper 1%	Upper 5%	Lower 5%	Lower 1%		Upper 1%	Upper 5%	Lower 5%	Lower 1%
50	4.88	3.99	2.15	1.95	600	3.54	3.34	2.70	2.60
75	4.59	3.87	2.27	2.08	650	3.52	3.33	2.71	2.61
100	4.39	3.77	2.35	2.18	700	3.50	3.31	2.72	2.62
125	4.24	3.71	2.40	2.24	750	3.48	3.30	2.73	2.64
150	4.13	3.65	2.45	2.29	800	3.46	3.29	2.74	2.65
					850	3.45	3.28	2.74	2.66
200	3.98	3.57	2.51	2.37	900	3.43	3.28	2.75	2.66
250	3.87	3.52	2.55	2.42	950	3.42	3.27	2.76	2.67
300	3.79	3.47	2.59	2.46	1000	3.41	3.26	2.76	2.68
350	3.72	3.44	2.62	2.50					
400	3.67	3.41	2.64	2.52	1200	3.37	3.24	2.78	2.71
450	3.63	3.39	2.66	2.55	1400	3.34	3.22	2.80	2.72
500	3.60	3.37	2.67	2.57	1600	3.32	3.21	2.81	2.74
550	3.57	3.35	2.69	2.58	1800	3.30	3.20	2.82	2.76
600	3.54	3.34	2.70	2.60	2000	3.28	3.18	2.83	2.77

Table A-2a Critical Values for the Kolmogorov-Smirnov Test of Goodness-of-Fit

Sample Size (n)	Significance Level .20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.829
4	.494	.525	.564	.624	.734
5	.446	.474	.510	.563	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.409	.486
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.391
17	.250	.266	.286	.318	.380
18	.244	.259	.278	.309	.370
19	.237	.252	.272	.301	.361
20	.231	.246	.264	.294	.352
25	.21	.22	.24	.264	.32
30	.19	.20	.22	.242	.29
35	.18	.19	.21	.23	.27
40				.21	.25
50				.19	.23
60				.17	.21
70				.16	.19
80				.15	.18
90				.14	
100				.14	
Asymptotic Formula:	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Table A-2b Critical Values for the Kolmogorov-Smirnov Test of Goodness-of-Fit (use sample mean and standard deviation)

Sample Size N	Level of Significance for $D = \text{Max} \lvert F^*(X) - S_N(X) \rvert$				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.149	.153	.165	.180	.203
30	.131	.136	.144	.161	.187
Over 30	$\frac{.736}{\sqrt{N}}$	$\frac{.768}{\sqrt{N}}$	$\frac{.805}{\sqrt{N}}$	$\frac{.886}{\sqrt{N}}$	$\frac{1.031}{\sqrt{N}}$

Table A-3 Cumulative Normal Distribution--Values of P

(Values of P corresponding to z_p for the normal curve. z is the standard normal variable. The value for P for $-z_p$ equals one minus the value of P for $+z_p$, e.g., the P for -1.62 equals 1 - .9474 = .0526.)

z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table A-4 Percentiles of the Student's t Distribution

ν \ q	60	75	90	95	97.5	99	99.5	99.95
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	.289	.816	1.886	2.920	4.303	6.965	9.925	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	12.941
4	.271	.741	1.533	2.132	2.776	3.747	4.604	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	6.859
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	5.405
8	.262	.706	2.397	1.860	2.306	2.896	3.355	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.105	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.291

Table A-5 Percentiles of the F-Distribution

95th Percentile

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	236·8	238·9	240·5	241·9	243·9	245·9	248·0	249·1	250·1	251·1	252·2	253·3	254·3
2	18·51	19·00	19·16	19·25	19·30	19·33	19·35	19·37	19·38	19·40	19·41	19·43	19·45	19·45	19·46	19·47	19·48	19·49	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·89	8·85	8·81	8·79	8·74	8·70	8·66	8·64	8·62	8·59	8·57	8·55	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·09	6·04	6·00	5·96	5·91	5·86	5·80	5·77	5·75	5·72	5·69	5·66	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·88	4·82	4·77	4·74	4·68	4·62	4·56	4·53	4·50	4·46	4·43	4·40	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·21	4·15	4·10	4·06	4·00	3·94	3·87	3·84	3·81	3·77	3·74	3·70	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·79	3·73	3·68	3·64	3·57	3·51	3·41	3·41	3·38	3·34	3·30	3·27	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·50	3·44	3·39	3·35	3·28	3·22	3·15	3·12	3·08	3·04	3·01	2·97	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·29	3·23	3·18	3·14	3·07	3·01	2·94	2·90	2·86	2·83	2·79	2·75	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·14	3·07	3·02	2·98	2·91	2·85	2·77	2·74	2·70	2·66	2·62	2·58	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	3·01	2·95	2·90	2·85	2·79	2·72	2·65	2·61	2·57	2·53	2·49	2·45	2·40
12	4·75	3·89	3·49	3·26	3·11	3·00	2·91	2·85	2·80	2·75	2·69	2·62	2·54	2·51	2·47	2·43	2·38	2·34	2·30
13	4·67	3·81	3·41	3·18	3·03	2·92	2·83	2·77	2·71	2·67	2·60	2·53	2·46	2·42	2·38	2·34	2·30	2·25	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·76	2·70	2·65	2·60	2·53	2·46	2·39	2·35	2·31	2·27	2·22	2·18	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·71	2·64	2·59	2·54	2·48	2·40	2·33	2·29	2·25	2·20	2·16	2·11	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·66	2·59	2·54	2·49	2·42	2·35	2·28	2·24	2·19	2·15	2·11	2·06	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·61	2·55	2·49	2·45	2·38	2·31	2·23	2·19	2·15	2·10	2·06	2·01	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·58	2·51	2·46	2·41	2·34	2·27	2·19	2·15	2·11	2·06	2·02	1·97	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·54	2·48	2·42	2·38	2·31	2·23	2·16	2·11	2·07	2·03	1·98	1·93	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·51	2·45	2·39	2·35	2·28	2·20	2·12	2·08	2·04	1·99	1·95	1·90	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·49	2·42	2·37	2·32	2·25	2·18	2·10	2·05	2·01	1·96	1·92	1·87	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·46	2·40	2·34	2·30	2·23	2·16	2·07	2·03	1·98	1·94	1·89	1·84	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·44	2·37	2·32	2·27	2·20	2·13	2·05	2·01	1·96	1·91	1·86	1·81	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·42	2·36	2·30	2·25	2·18	2·11	2·03	1·98	1·94	1·89	1·84	1·79	1·73
25	4·24	3·39	2·99	2·76	2·60	2·49	2·40	2·34	2·28	2·24	2·16	2·09	2·01	1·96	1·92	1·87	1·82	1·77	1·71
26	4·23	3·37	2·98	2·74	2·59	2·47	2·39	2·32	2·27	2·22	2·15	2·07	1·99	1·95	1·90	1·85	1·80	1·75	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·37	2·31	2·25	2·20	2·13	2·06	1·97	1·93	1·88	1·84	1·79	1·73	1·67
28	4·20	3·34	2·95	2·71	2·56	2·45	2·36	2·29	2·24	2·19	2·12	2·04	1·96	1·91	1·87	1·82	1·77	1·71	1·65
29	4·18	3·33	2·93	2·70	2·55	2·43	2·35	2·28	2·22	2·18	2·10	2·03	1·94	1·90	1·85	1·81	1·75	1·70	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·33	2·27	2·21	2·16	2·09	2·01	1·93	1·89	1·84	1·79	1·74	1·68	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·25	2·18	2·12	2·08	2·00	1·92	1·84	1·79	1·74	1·69	1·64	1·58	1·51
60	4·00	3·15	2·76	2·53	2·37	2·25	2·17	2·10	2·04	1·99	1·92	1·84	1·75	1·70	1·65	1·59	1·53	1·47	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·09	2·02	1·96	1·91	1·83	1·75	1·66	1·61	1·55	1·50	1·43	1·35	1·26
∞	3·84	3·00	2·60	2·37	2·21	2·10	2·01	1·94	1·88	1·83	1·75	1·67	1·57	1·52	1·46	1·39	1·32	1·22	1·00

Table A-5 - Cont'd

99th Percentile

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999·5	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98·50	99·00	99·17	99·25	99·30	99·33	99·36	99·37	99·39	99·40	99·42	99·43	99·45	99·46	99·47	99·47	99·48	99·49	99·50
3	34·12	30·82	29·46	28·71	28·24	27·91	27·67	27·49	27·35	27·23	27·05	26·87	26·69	26·60	26·50	26·41	26·32	26·22	26·13
4	21·20	18·00	16·69	15·98	15·52	15·21	14·98	14·80	14·66	14·55	14·37	14·20	14·02	13·93	13·84	13·75	13·65	13·56	13·46
5	16·26	13·27	12·06	11·39	10·97	10·67	10·46	10·29	10·16	10·05	9·89	9·72	9·55	9·47	9·38	9·29	9·20	9·11	9·02
6	13·75	10·92	9·78	9·15	8·75	8·47	8·26	8·10	7·98	7·87	7·72	7·56	7·40	7·31	7·23	7·14	7·06	6·97	6·88
7	12·25	9·55	8·45	7·85	7·46	7·19	6·99	6·84	6·72	6·62	6·47	6·31	6·16	6·07	5·99	5·91	5·82	5·74	5·65
8	11·26	8·65	7·59	7·01	6·63	6·37	6·18	6·03	5·91	5·81	5·67	5·52	5·36	5·28	5·20	5·12	5·03	4·95	4·86
9	10·56	8·02	6·99	6·42	6·06	5·80	5·61	5·47	5·35	5·26	5·11	4·96	4·81	4·73	4·65	4·57	4·48	4·40	4·31
10	10·04	7·56	6·55	5·99	5·64	5·39	5·20	5·06	4·94	4·85	4·71	4·56	4·41	4·33	4·25	4·17	4·08	4·00	3·91
11	9·65	7·21	6·22	5·67	5·32	5·07	4·89	4·74	4·63	4·54	4·40	4·25	4·10	4·02	3·94	3·86	3·78	3·69	3·60
12	9·33	6·93	5·95	5·41	5·06	4·82	4·64	4·50	4·39	4·30	4·16	4·01	3·86	3·78	3·70	3·62	3·54	3·45	3·36
13	9·07	6·70	5·74	5·21	4·86	4·62	4·44	4·30	4·19	4·10	3·96	3·82	3·66	3·59	3·51	3·43	3·34	3·25	3·17
14	8·86	6·51	5·56	5·04	4·69	4·46	4·28	4·14	4·03	3·94	3·80	3·66	3·51	3·43	3·35	3·27	3·18	3·09	3·00
15	8·68	6·36	5·42	4·89	4·56	4·32	4·14	4·00	3·89	3·80	3·67	3·52	3·37	3·29	3·21	3·13	3·05	2·96	2·87
16	8·53	6·23	5·29	4·77	4·44	4·20	4·03	3·89	3·78	3·69	3·55	3·41	3·26	3·18	3·10	3·02	2·93	2·84	2·75
17	8·40	6·11	5·18	4·67	4·34	4·10	3·93	3·79	3·68	3·59	3·46	3·31	3·16	3·08	3·00	2·92	2·83	2·75	2·65
18	8·29	6·01	5·09	4·58	4·25	4·01	3·84	3·71	3·60	3·51	3·37	3·23	3·08	3·00	2·92	2·84	2·75	2·66	2·57
19	8·18	5·93	5·01	4·50	4·17	3·94	3·77	3·63	3·52	3·43	3·30	3·15	3·00	2·92	2·84	2·76	2·67	2·58	2·49
20	8·10	5·85	4·94	4·43	4·10	3·87	3·70	3·56	3·46	3·37	3·23	3·09	2·94	2·86	2·78	2·69	2·61	2·52	2·42
21	8·02	5·78	4·87	4·37	4·04	3·81	3·64	3·51	3·40	3·31	3·17	3·03	2·88	2·80	2·72	2·64	2·55	2·46	2·36
22	7·95	5·72	4·82	4·31	3·99	3·76	3·59	3·45	3·35	3·26	3·12	2·98	2·83	2·75	2·67	2·58	2·50	2·40	2·31
23	7·88	5·66	4·76	4·26	3·94	3·71	3·54	3·41	3·30	3·21	3·07	2·93	2·78	2·70	2·62	2·54	2·45	2·35	2·26
24	7·82	5·61	4·72	4·22	3·90	3·67	3·50	3·36	3·26	3·17	3·03	2·89	2·74	2·66	2·58	2·49	2·40	2·31	2·21
25	7·77	5·57	4·68	4·18	3·85	3·63	3·46	3·32	3·22	3·13	2·99	2·85	2·70	2·62	2·54	2·45	2·36	2·27	2·17
26	7·72	5·53	4·64	4·14	3·82	3·59	3·42	3·29	3·18	3·09	2·96	2·81	2·66	2·58	2·50	2·42	2·33	2·23	2·13
27	7·68	5·49	4·60	4·11	3·78	3·56	3·39	3·26	3·15	3·06	2·93	2·78	2·63	2·55	2·47	2·38	2·29	2·20	2·10
28	7·64	5·45	4·57	4·07	3·75	3·53	3·36	3·23	3·12	3·03	2·90	2·75	2·60	2·52	2·44	2·35	2·26	2·17	2·06
29	7·60	5·42	4·54	4·04	3·73	3·50	3·33	3·20	3·09	3·00	2·87	2·73	2·57	2·49	2·41	2·33	2·23	2·14	2·03
30	7·56	5·39	4·51	4·02	3·70	3·47	3·30	3·17	3·07	2·98	2·84	2·70	2·55	2·47	2·39	2·30	2·21	2·11	2·01
40	7·31	5·18	4·31	3·83	3·51	3·29	3·12	2·99	2·89	2·80	2·66	2·52	2·37	2·29	2·20	2·11	2·02	1·92	1·80
60	7·08	4·98	4·13	3·65	3·34	3·12	2·95	2·82	2·72	2·63	2·50	2·35	2·20	2·12	2·03	1·94	1·84	1·73	1·60
120	6·85	4·79	3·95	3·48	3·17	2·96	2·79	2·66	2·56	2·47	2·34	2·19	2·03	1·95	1·86	1·76	1·66	1·53	1·38
∞	6·63	4·61	3·78	3·32	3·02	2·80	2·64	2·51	2·41	2·32	2·18	2·04	1·88	1·79	1·70	1·59	1·47	1·32	1·00

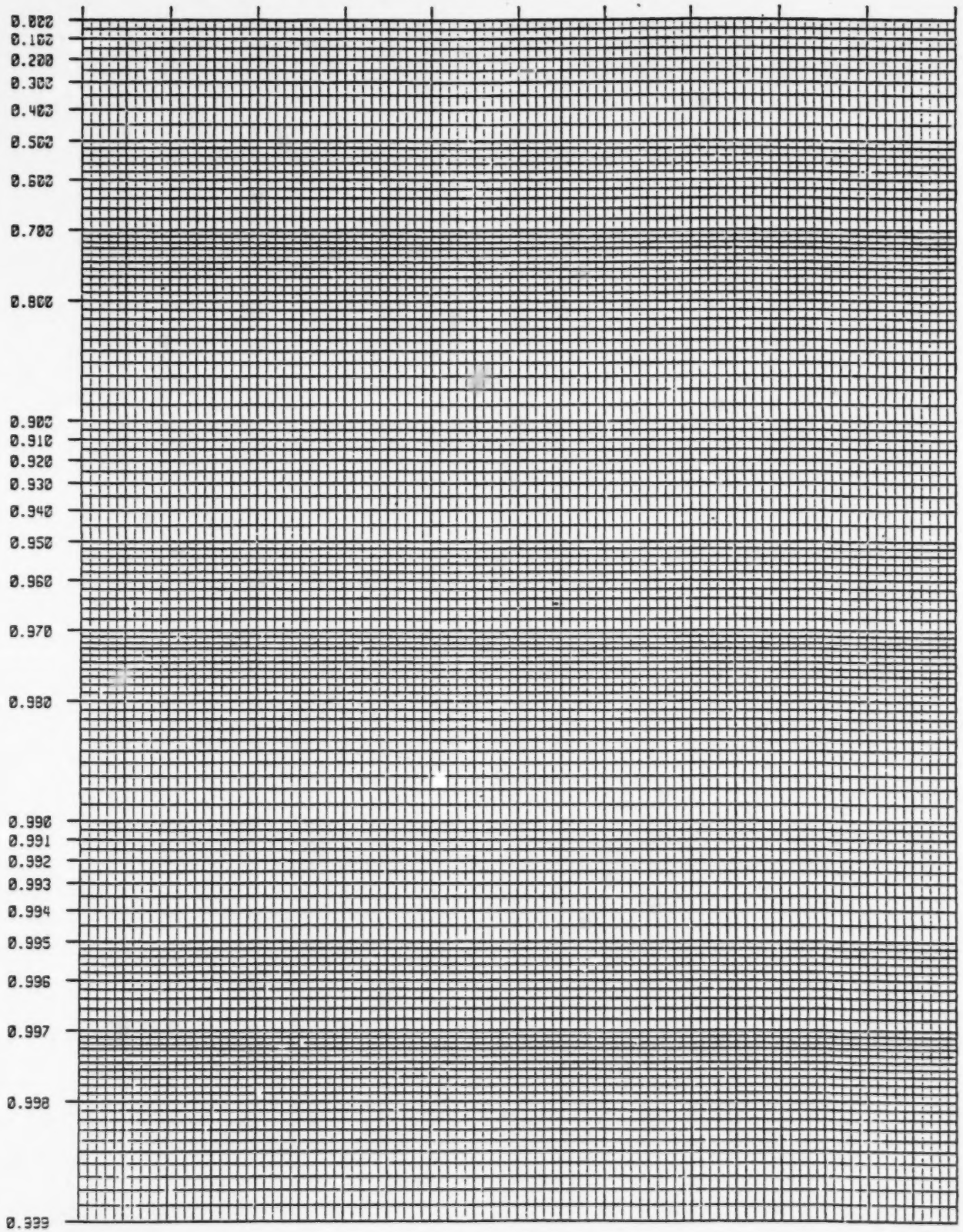
Table A-6 Percentiles of the χ^2 Distribution

ν \ q	0.5	1	2.5	5	10	20	30	40
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	0.0642	0.148	0.275
2	0.0100	0.0201	0.0506	0.103	0.211	0.446	0.713	1.02
3	0.0717	0.115	0.216	0.352	0.584	1.00	1.42	1.87
4	0.207	0.297	0.484	0.711	1.06	1.65	2.19	2.75
5	0.412	0.554	0.831	1.15	1.61	2.34	3.00	3.66
6	0.676	0.872	1.24	1.64	2.20	3.07	3.83	4.57
7	0.989	1.24	1.69	2.17	2.83	3.82	4.67	5.49
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	6.42
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	7.36
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	8.30
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	9.24
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	10.2
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	11.1
14	4.07	4.66	5.63	6.57	7.79	9.47	10.8	12.1
15	4.60	5.23	6.26	7.26	8.55	10.3	11.7	13.0
16	5.14	5.81	6.91	7.96	9.31	11.2	12.6	14.0
17	5.70	6.41	7.56	8.67	10.1	12.0	13.5	14.9
18	6.26	7.01	8.23	9.39	10.9	12.9	14.4	15.9
19	6.84	7.63	8.91	10.1	11.7	13.7	15.4	16.9
20	7.43	8.26	9.59	10.9	12.4	14.6	16.3	17.8
21	8.03	8.90	10.3	11.6	13.2	15.4	17.2	18.8
22	8.64	9.54	11.0	12.3	14.0	16.3	18.1	19.7
23	9.26	10.2	11.7	13.1	14.8	17.2	19.0	20.7
24	9.89	10.9	12.4	13.8	15.7	18.1	19.9	21.7
25	10.5	11.5	13.1	14.6	16.5	18.9	20.9	22.6
26	11.2	12.2	13.8	15.4	17.3	19.8	21.8	23.6
27	11.8	12.9	14.6	16.2	18.1	20.7	22.7	24.5
28	12.5	13.6	15.3	16.9	18.9	21.6	23.6	25.5
29	13.1	14.3	16.0	17.7	19.8	22.5	24.6	26.5
30	13.8	15.0	16.8	18.5	20.6	23.4	25.5	27.4
35	17.2	18.5	20.6	22.5	24.8	27.8	30.2	32.3
40	20.7	22.2	24.4	26.5	29.1	32.3	34.9	37.1
45	24.3	25.9	28.4	30.6	33.4	36.9	39.6	42.0
50	28.0	29.7	32.4	34.8	37.7	41.4	44.3	46.9
75	47.2	49.5	52.9	56.1	59.8	64.5	68.1	71.3
100	67.3	70.1	74.2	77.9	82.4	87.9	92.1	95.8

Table A-6 - Cont'd

ν \ q	50	60	70	80	90	95	97.5	99	99.5	99.9
1	0.455	0.708	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.8
2	1.39	1.83	2.41	3.22	4.61	5.99	7.38	9.21	10.6	13.8
3	2.37	2.95	3.67	4.64	6.25	7.81	9.35	11.3	12.8	16.3
4	3.36	4.04	4.88	5.99	7.78	9.49	11.1	13.3	14.9	18.5
5	4.35	5.13	6.06	7.29	9.24	11.1	12.8	15.1	16.7	20.5
6	5.35	6.21	7.23	8.56	10.6	12.6	14.4	16.8	18.5	22.5
7	6.35	7.28	8.38	9.80	12.0	14.1	16.0	18.5	20.3	24.3
8	7.34	8.35	9.52	11.0	13.4	15.5	17.5	20.1	22.0	26.1
9	8.34	9.41	10.7	12.2	14.7	16.9	19.0	21.7	23.6	27.9
10	9.34	10.5	11.8	13.4	16.0	18.3	20.5	23.2	25.2	29.6
11	10.3	11.5	12.9	14.6	17.3	19.7	21.9	24.7	26.8	31.3
12	11.3	12.6	14.0	15.8	18.5	21.0	23.3	26.2	28.3	32.9
13	12.3	13.6	15.1	17.0	19.8	22.4	24.7	27.7	29.8	34.5
14	13.3	14.7	16.2	18.2	21.1	23.7	26.1	29.1	31.3	36.1
15	14.3	15.7	17.3	19.3	22.3	25.0	27.5	30.6	32.8	37.7
16	15.3	16.8	18.4	20.5	23.5	26.3	28.8	32.0	34.3	39.3
17	16.3	17.8	19.5	21.6	24.8	27.6	30.2	33.4	35.7	40.8
18	17.3	18.9	20.6	22.8	26.0	28.9	31.5	34.8	37.2	42.3
19	18.3	19.9	21.7	23.9	27.2	30.1	32.9	36.2	38.6	43.8
20	19.3	21.0	22.8	25.0	28.4	31.4	34.2	37.6	40.0	45.3
21	20.3	22.0	23.9	26.9	29.6	32.7	35.5	38.9	41.4	46.8
22	21.3	23.0	24.9	27.3	30.8	33.9	36.8	40.3	42.8	48.3
23	22.3	24.1	26.0	28.4	32.0	35.2	38.1	41.6	44.2	49.7
24	23.3	25.1	27.1	29.6	33.2	36.4	39.4	43.0	45.6	51.2
25	24.3	26.1	28.2	30.7	34.4	37.7	40.6	44.3	46.9	52.6
26	25.3	27.2	29.2	31.8	35.6	38.9	41.9	45.6	48.3	54.1
27	26.3	28.2	30.3	32.9	36.7	40.1	43.2	47.0	49.6	55.5
28	27.3	29.2	31.4	34.0	37.9	41.3	44.5	48.3	51.0	56.9
29	28.3	30.3	32.5	35.1	39.1	42.6	45.7	49.6	52.3	58.3
30	29.3	31.3	33.5	36.3	40.3	43.8	47.0	50.9	53.7	59.7
35	34.3	36.5	38.9	41.8	46.1	49.8	53.2	57.3	60.3	66.6
40	39.3	41.6	44.2	47.3	51.8	55.8	59.3	63.7	66.8	73.4
45	44.3	46.8	49.5	52.7	57.5	61.7	65.4	70.0	73.2	80.1
50	49.3	51.9	54.7	58.2	63.2	67.5	71.4	76.2	79.5	86.7
75	74.3	77.5	80.9	85.1	91.1	96.2	100.8	106.4	110.3	118.6
100	99.3	102.9	106.9	111.7	118.5	124.3	129.6	135.6	140.2	149.4

Table A-7 Sample of Semilog Probability Paper



APPENDIX B

TEST FOR BIVARIATE NORMALITY

The following artical presents a method for testing the hypothese that two factors of an EFE are bivariate normal.

STANDARD TEST FOR BIVARIATE NORMALITY*

2.1.2 Stage 2. Assessment of the Distributional Properties of the Emission Data

An assumption of multivariate normality underlies much of the standard, classical multivariate statistical methodology. But in real data, possibility of departure from normality is prevalent. Thus, it is essential to check the normality assumption before applying the classical theory. Such a check will be helpful not only in guiding the subsequent analysis, but also in suggesting the need for a transformation

*From Approaches to Analyzing Data from the Portland Study, TSC-PD-A158-5, Final Report on Task 4, Contract No. 68-03-2392, May 1979, pp. 2-14 through 2-35.

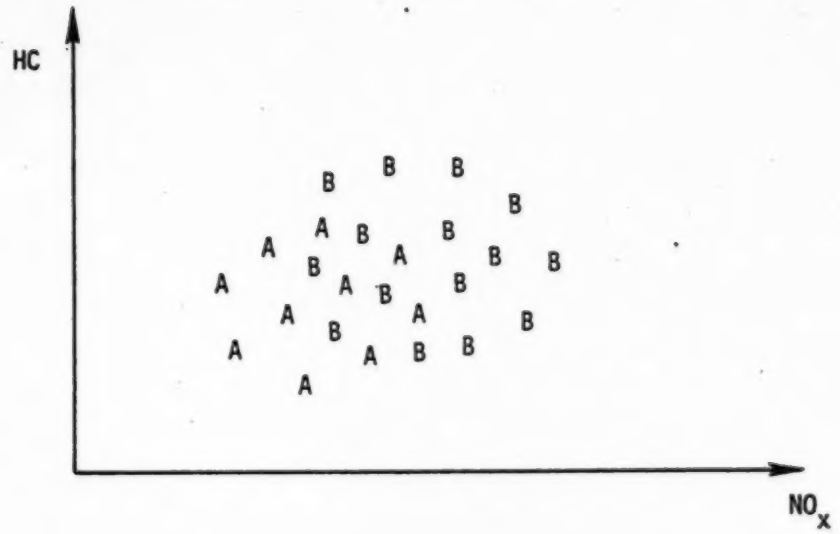


Figure 2.6 Scatterplot, with points from Categories I and II identified by distinct letters.

of the data to make them more nearly normally distributed (and therefore capable of having the classical theory applied).

The joint and marginal distributions of the emissions measured by the short test for in-use vehicles have not been checked; such a check will be necessary once the Portland data are made available.

The following steps for assessing joint normality are recommended:

Step 1. Test for marginal normality (FTP and ST separately).

Step 2. If the hypothesis of normality is rejected in Step 1, perform a transformation of the data and redo Step 1. (NOTE: If it is obvious that the data cannot be normalized, go to Stage 3.)

Step 3. Test for joint normality.

Step 1. Test for Marginal Normality. Although marginal normality does not imply joint normality, the presence of many types of non-normality is often reflected in the marginal distributions as well. Hence, a natural, simple, and preliminary step in evaluating the normality of multivariate data is to study the reasonableness of marginal normality for the observations on each of the variables.

Two commonly used techniques for assessing marginal normalities are:

- normal probability plot
- skewness and kurtosis coefficients and the Kolmogorov-Smirnov (K-S) test.

A normal probability plot is a graphical tool for assessing univariate normality. The technique consists of

1. Ordering the n observations from smallest to largest, where $X_{(i)}$ is the i^{th} smallest observation in the sample, i=1,...,n, and $X_{(1)} < X_{(2)} <...< X_{(n)}$.
2. Letting $P_i=(i - 1/2)/n$.
3. Plotting $X_{(i)}$ against P_i on normal probability paper.

A linear configuration on such a plot would correspond to adequate normality of the observations (Figure 2.7a), while systematic and subtle departure from normality would be indicated by deviations from linearity (Figure 2.7b).

Computing Support: TSC computer library, BMDP5D (the histogram program mentioned in Stage 1).

Skewness and kurtosis coefficients and the K-S test provide a statistics-based formal test on univariate normality. Let $X_1,..., X_n$ be n univariate emission observations. The skewness and kurtosis coefficients are defined as

$$\sqrt{b_1} = \frac{\sqrt{n} \sum_{i=1}^{n} (X_i - \bar{X})^3}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]^{3/2}}$$

$$b_2 = \frac{n \sum_{i=1}^{n} (X_i - \bar{X})^4}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]^2} ,$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$.

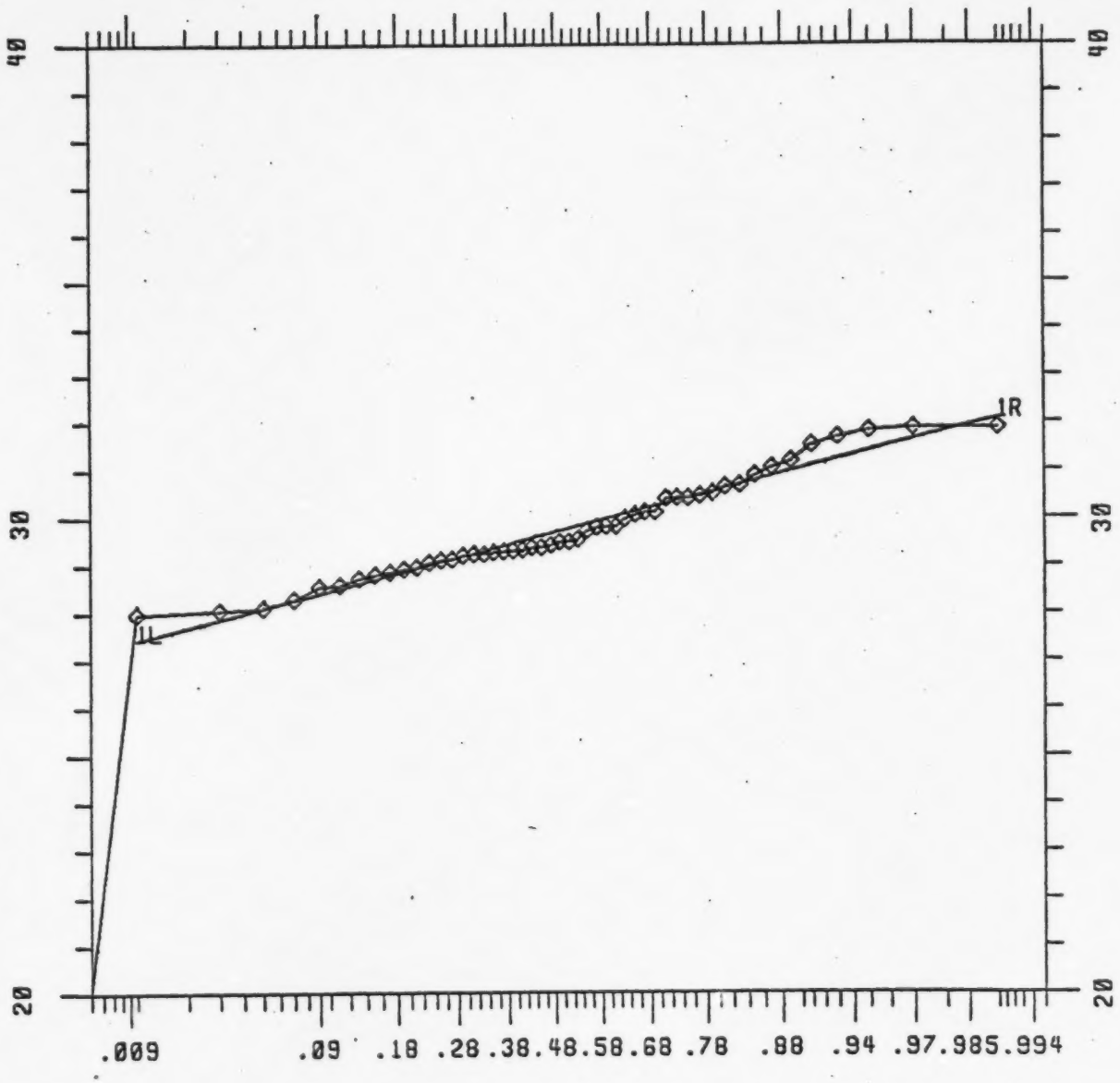


Figure 2.7a Simulated normal data plotted on normal paper.

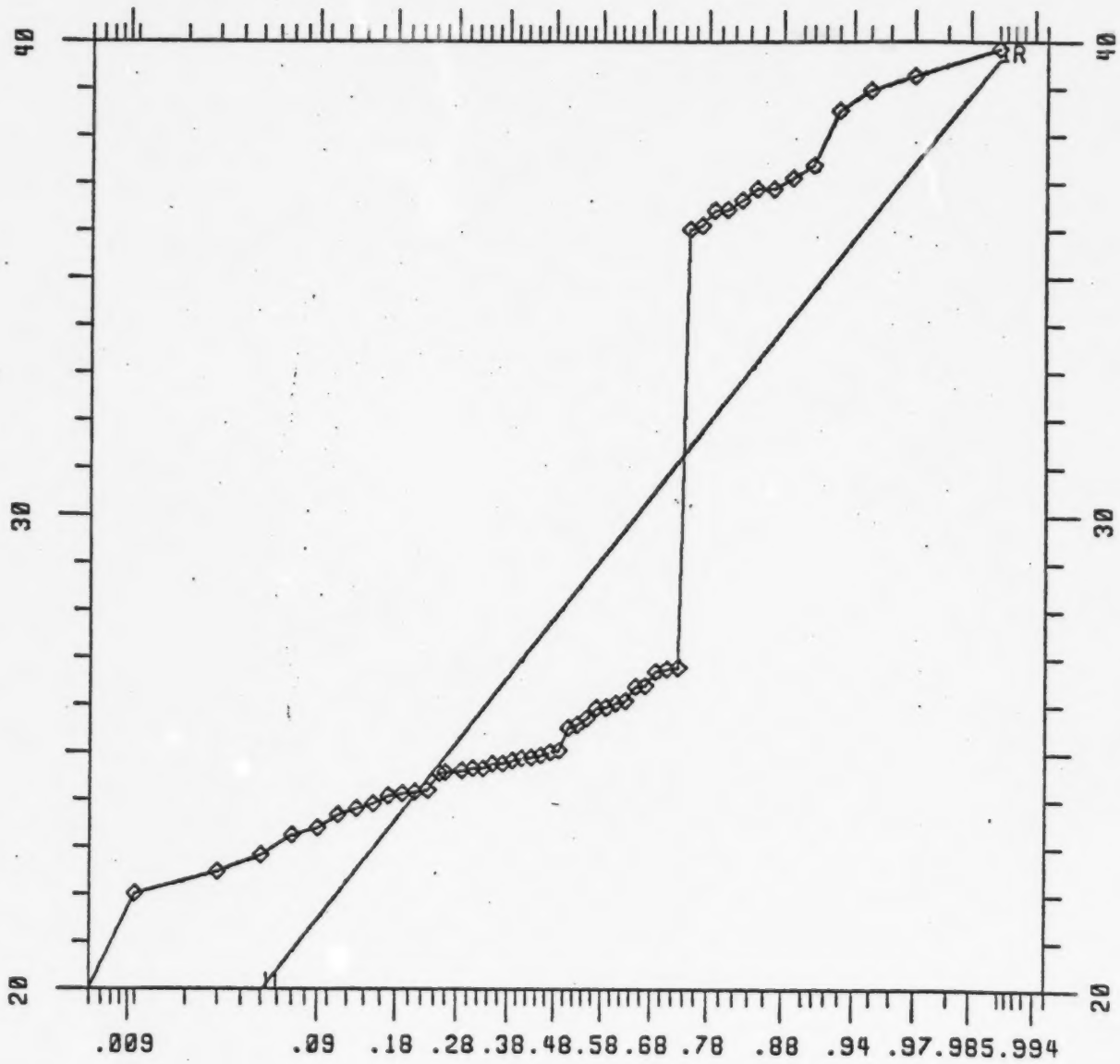


Figure 2.7b Simulated non-normal data plotted on normal paper.

The following are characteristics of $\sqrt{b_1}$ and b_2:

1. When the population density is symmetric, then $b_1=0$. When the population density has a long tail to the right, then $\sqrt{b_1} > 0$; and when it has a long tail to the left, $\sqrt{b_1} < 0$.
2. The value of b_2 for any normal distribution is $b_2=3$. But if the distribution is concentrated toward the mean, then $b_2 < 3$; otherwise, $b_2 > 3$.

For real data, the values of these coefficients are subject to random fluctuations; therefore, a statistical table has to be used to determine the value. Tables of approximately 5% and 1% points of these two statistics are given in Appendix A. (Table A-1). The procedures for testing skewness and kurtosis are given below.

SKEWNESS & KURTOSIS TEST ($\sqrt{b_1}=0$ and $b_2=3$)

PROCEDURE	EXAMPLE
(1) Select a variable to be tested, say, variable X. (This may be the original variable or the transformed variable.)	(1) Suppose, from a histogram plot in Stage 1, that the CO measured from ST for all the cars passing the FTP looks normal. One then decides to test the ST-CO variable for normality.
(2) Set a significance level, α. Following the convention, set α equal to either 0.01 or 0.05.	(2) Set $\alpha=0.05$

(3) Compute the skewness coefficient, $\sqrt{b_1}$, and kurtosis coefficient, b_2, for the variable X. These two numbers can be easily obtained from the output of a frequency count subroutine such as BMDP2D or SPSS-FREQUENCY SUBPROGRAM.

(3) Using BMDP2D subroutine, one obtains $\sqrt{b_1}=0.2$ and $b_2=3.5$.

(4) Let N be the number of sample points. Look for the one-tailed percentage point, p_1, of the distribution of $\sqrt{b_1}$ determined by N and α from the table for testing skewness (see Table A-1a). Also, look for percentage point p_2 of the distribution of b_2 determined by N and α from the table for testing kurtosis (see Table A-1b).

(4) Suppose there are 112 sample points. From the table for testing skewness, one obtains the one-tailed percentage point of the distribution of $\sqrt{b_1}$ as $p_1=0.350$ (determined by 112 and 0.05). Similarly, from the table for testing kurtosis, one obtains the percentage point of the distribution of b_2 or $p_2=3.71$.

(5) If $\sqrt{b_1}$ obtained from (3) is smaller than p_1, accept the hypothesis that $\sqrt{b_1}=0$ at α significance level and thus passes the skewness test for normality. Otherwise, reject the hypothesis that $\sqrt{b_1}=0$;

(5) Since $\sqrt{b_1}=0.2 < 0.35$, it follows that the hypothesis $\sqrt{b_1}=0$ is accepted at 0.05 significance level. Also from $b_2=3.5 <3.71$, it follows that the hypothesis $b_2=3$ is accepted at 0.05 significance level. Therefore,

X, therefore is not normal. Similarly, if b_2 obtained from (3) is smaller than p_2, accept the hypothesis that $b_2=3$ at α significance level and thus passes the kurtosis test for normality. Otherwise, reject the hypothesis that $b_2=3$; X, therefore, is not normal.

X passes the first test of normality and thus can proceed for a further check for normality as shown in the next procedure.

In the event that $\sqrt{b_1}=0$ and $b_2=3$ have been statistically accepted, which is an indication of normality, the K-S test can be utilized to further test the hypothesis that the observations are normally distributed. The procedure for the K-S test can be described as follows:

KOLMOGOROV-SMIRNOV TEST

PROCEDURE

(1) Select the variable to be tested, say, X. (This may be the original variable or the transformed variable).

(2) Set a significance level α. Following the convention, set $\alpha=0.01$ or 0.05.

EXAMPLE

(1) Suppose, in Stage 1, that the histogram of CO measured from ST for cars passing the FTP looks like normal. Then one decides to formally test the normality of the ST-CO variable for cars passing the FTP.

(2) Set $\alpha=0.05$

(3) Calculate the sample mean, $\bar{X}$, and sample variance S^2 of variable X. These two numbers can be easily obtained from the output of a descriptive subroutine such as BMDP5D or SPSS-DESCRIPTIVE SUBPROGRAM.

(3) Using SPSS-DESCRIPTIVE SUBPROGRAM, one obtains the sample mean $\bar{X}$, and sample variance, S^2, of FTP-CO variable as $\bar{X}$=34.5 and S^2=2.0.

(4) Set the null hypothesis, H_0: X follows a normal distribution with mean $\bar{X}$ and variance S^2, which can be written ($N(\bar{X},S^2)$). Implement the subroutine KOLMO of the SSP-1 (available through MIDAS and the TSC computing library) on X data with hypothesized distribution as stated in H_0.

(4) Set the null hypothesis as H_0: ST-CO follows a normal distribution with mean 34.5 and variance 2.0. Implement the subroutine KOLMO of the SSP-1 on ST-CO data with hypothesized distribution as stated in H_0.

(5) From the printout of the KOLMO program, look for the K-S distance, D. For the significance α-level, look up the critical value P for the K-S test of goodness-of-fit from Table A-2b (see Appendix A). (NOTE: If the number of sample points is less than 100, find the critical value from the table under the α-level column. If the number of sample points

(5) From the printout of the KOLMO program, look for the K-S distance D and find D=0.035. For the 0.05 significance level, look up the critical value P from Table A-2b. Since there are 112 sample points, the asymptotic formula for obtaining P is used and we get $P=0.886/\sqrt{(112)}=0.083$. Since P>D, accept the hypothesis H_0, i.e., X follows N(34.5,2.0).

is greater than 100, compute the critical value from the asymptotic formula.) If $D<P$, then accept the hypothesis H_0, i.e., X follows $N(\bar{X},S^2)$. Otherwise, reject the hypothesis, i.e., X does not follow the normal distribution.

Step 2. Transformation of the Data. If, for a particular emission variable, the hypothesis of normality is rejected at Step 1, then a transformation of the variable that induces normality should be considered. Logarithmic transformations and square-root transformations are the two commonly used types.

EXAMPLE: The histogram plot of the HC values measured from the short test is highly skewed with a long tail to the right (Figure 2.5b, p. 2-12); thus, the logarithms of individual HC observations are utilized, and the resultant transformed sample histogram now looks normal. Going again through Step 1, the hypothesis of normality is accepted. Therefore, the HC variable follows a lognormal distribution. Alternatively, the square-root transformation can be used, and the acceptance of the normality of the transformed variable will mean the original variable follows a chi-square distribution.

Step 3. Test for Joint Normality. After marginal normality is confirmed for all the individual emission variables (or transformed variables), the test for joint normality can be implemented for a final check.

Since there are only three emission variables, it usually suffices to test the bivariate normality for each pair of variables.

Two methods for assessing the joint normality are:

- A test based on linear regressions
- A test based on radius-and-angle representation of multivariate data.

A property of the multivariate normal distribution is that the regression of each variable on all others is linear. Thus, a simple test for detecting joint non-normality would be to perform a linear regression line fit and decide from the residual plot whether the regression shows any sign of curvature. Figure 2.8 illustrates two possible situations. If the residual plot indicates obvious curvature, reject the hypothesis of joint normality; otherwise, accept bivariate normality.

Computing Support: Any linear regression program with a residual plot will do.

The second procedure is based on two things: 1) representing a two-dimensional data point in terms of radius and angle; and 2) under the joint normal assumption, the radius's following a chi-squared distribution with 2 degrees of freedom and the angle's following a uniform distribution over $(-\pi/2,\pi/2)$. Specifically, the operations of this procedure can be described as follows:

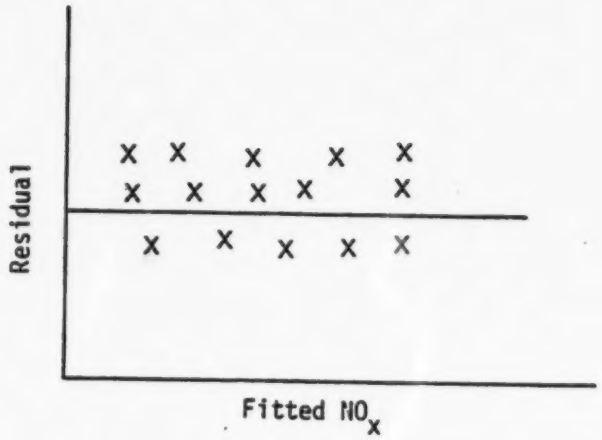


Figure 2.8a Case for the acceptance for possible joint normality of NO_x and CO.

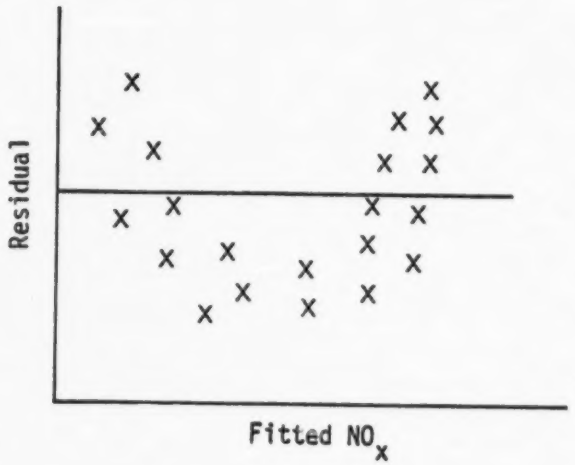


Figure 2.8b Residual plot indicating curvature; the hypothesis of joint normality is rejected.

Figure 2.8 Linear regression line fit for assessing the joint normality of NO_x and CO, where $\widehat{NO}_x = a + b(CO)$ and the residual = $NO_x - \widehat{NO}_x$.

PROCEDURE

(1) Let $\underline{X}_1, \ldots \underline{X}_n$ be n two-dimensional observations $\underline{X}_k = (X_{1k}, X_{2k})$. Compute the covariance matrix S where

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{pmatrix}$$

$$S_{ij} = \frac{1}{n-1}\left[\left(\sum_{k=1}^{n} X_{ik}\, X_{jk}\right) - \left(\sum_{k=1}^{n} X_{ik}\right)\left(\sum_{k=1}^{n} X_{jk}\right)\Big/ n\right]$$

and the mean vector $\overline{\underline{X}}$

$\overline{\underline{X}} = (\overline{X}_1, \overline{X}_2)$, where

$$\overline{X}_i = \frac{\sum_{k=1}^{n} X_{ik}}{n} .$$

(2) Find θ_k and r_k^2 for each two-dimensional observation.

where $\theta_k = \tan^{-1}\left(\frac{Z_{2i}}{Z_{1i}}\right)$

(remember to account for the correct quadrant)

and $r_k^2 = Z_{1k}^2 + Z_{2k}^2$, where

EXAMPLE

(1) From Table 2.1, which lists HC and NO_2 data,

$n = 47$

$S_{11} = 1.94570$

$S_{22} = 8.52141$

$S_{12} = 0.6301$

$\overline{X}_1 = 2.84383$

$\overline{X}_2 = 3.76447$

(2) Transform the data

$$Z_{1k} = \frac{(X_{1k}-2.84383)8.5214-(X_{2k}-3.76447)0.6301}{11.74320}$$

$$Z_{2k} = \frac{(X_{2k}-3.76447)}{2.91915}$$

The values appear in Table 2.1.

$$Z_{1k} = \frac{(X_{1k}-\bar{X}_1)S_{22} - (X_{2k}-\bar{X}_2)S_{12}}{\sqrt{S_{22}}\sqrt{S_{11}\,S_{22} - S_{12}^2}}$$

$$Z_{2k} = \frac{(X_{2k} - \bar{X}_2)}{\sqrt{S_{22}}}$$

(3) Compute SR and SLR

$$SR = \sum_{k=1}^{n} r_k^2 \text{ and}$$

$$SLR = \sum_{k=1}^{n} r_k^2 \log_e(r_k^2)$$

Calculate the goodness-of-fit statistic G

$$G = \left| \frac{\sqrt{n}}{0.887} \left(\frac{SLR}{SR} - \log_e(SR) + \log_e(n) - 0.4228 \right) \right| .$$

G has a standard normal distribution.

(4) Sort the θ values and transform them to the (0,1) interval. The transformation is:

$$0.5 + \frac{\theta_k}{2\pi} .$$

Compute the cumulative frequency distribution of the k^{th} ordered variable

$$\frac{k}{n+1} .$$

(3) Compute SR, SLR and G

SR = 243.786

SLR = 874.178

G = 1.517

(4) Table 2.2 gives the sorted $0.5 + \frac{\theta_k}{2\pi}$ values, $\frac{k}{n+1}$ and KS_k.

KS = 0.34862

Calculate the Kolmogorov-Smirnov distance for each value

$$KS_k = \left| 0.5 + \frac{\theta_k}{2\pi} - \frac{k}{n+1} \right|$$

and let KS be the maximum, i.e.,

$$KS = \max_k \left| 0.5 + \frac{\theta_k}{2\pi} - \frac{k}{n+1} \right| .$$

(5) Check G and KS to see if there is a significant difference.

(5) G = 1.517 and KS = 0.34862 from the normal table (Table A-3). One finds that at 0.05 significance level, G is insignificant, i.e., the distribution of r^2 is χ^2. However, from the K-S two-sample table (Table A-2a), one finds that at 0.05 significance level, KS is significant, i.e., the hypothesis that the distribution of θ is uniformly distributed is rejected. Therefore, the data is not bivariate normal. Transformation of the variable may be needed.

(6) Plot r(i) against $1-P_i$ on semilog probability paper, Figure 2.9a (the actual paper is presented as Table A-7) and plot $\theta_{(i)}$ against P_i on uniform probability paper, Figure 2.9b. If the data conforms statistically to the null hypothesis of bivariate normality, the configuration on these two probability plots should be reasonably linear. Departure from linearity on either or both of the plots would indicate specific types of departure from the null hypothesis.

Computing Support: TSC computer library.

TABLE 2.1 BIVARIATE HC AND NOX VALUES WITH ASSOCIATED Z_1, Z_2, θ, r^2 AND $r^2 \log_e r^2$

HC	NOX	Z_1	Z_2	θ	r^2	$r^2 \log_e r^2$
4.8	4.44	1.38324	0.23141	0.16576	1.96691	1.33054
2.64	7.00	-0.32152	1.10838	1.85312	1.33188	0.38171
6.35	1.84	2.64750	-0.65926	-0.24405	7.44388	14.94279
2.53	2.57	-0.16364	-0.40918	-1.95123	0.19421	-0.31827
2.02	2.20	-0.51387	-0.53593	-2.33517	0.55120	-0.32829
2.68	1.82	-0.01455	-0.66611	-1.59263	0.44391	-0.36051
15.85	1.02	9.56514	-0.94016	-0.09777	92.75883	420.19779
2.52	1.49	-0.11295	-0.77916	-1.71475	0.61984	-0.29647
1.82	1.12	-0.60105	-0.90591	-2.15658	1.18192	0.19755
11.66	4.57	6.35420	0.27595	0.04340	40.45202	149.67719
14.15	4.84	8.14657	0.36844	0.04520	66.50241	279.12646
3.45	1.69	0.55117	-0.71064	-0.91111	0.80881	-0.17163
3.31	1.57	0.45602	-0.75175	-1.02553	0.77309	-0.19897
3.30	3.01	0.37150	-0.25846	-0.60784	0.20481	-0.32476
2.45	3.64	-0.27910	-0.04264	-2.98999	0.07972	-0.20163
4.50	5.30	1.11940	0.52602	0.43929	1.52976	0.65032
3.43	1.44	0.55008	-0.79628	-0.96627	0.93665	-0.06130
2.76	2.51	0.00648	-0.42974	-1.55572	0.18472	-0.31197
1.15	2.25	-1.14786	-0.51881	-2.71709	1.58674	0.73257
3.48	3.21	0.49139	-0.18994	-0.36885	0.27754	-0.35575
3.57	6.30	0.39089	0.86859	1.14791	0.90724	-0.08832
3.46	5.55	0.35132	0.61166	1.04944	0.49755	-0.34732
1.26	2.33	-1.07233	-0.49140	-2.71189	1.39137	0.45955
3.01	2.70	0.17770	-0.36465	-1.11736	0.16455	-0.29693
1.97	3.46	-0.61775	-0.10430	-2.97433	0.39250	-0.36707
3.26	2.33	0.37896	-0.49140	-0.91387	0.38509	-0.36748
3.62	1.77	0.67024	-0.68324	-0.79500	0.91604	-0.08033
1.05	1.80	-1.19628	-0.67296	-2.62917	1.88396	1.19326
2.20	1.47	-0.34408	-0.78601	-1.98342	0.73620	-0.22546
1.63	1.61	-0.76521	-0.73805	-2.37426	1.13026	0.13840
1.61	1.75	-0.78723	-0.69009	-2.42186	1.09596	0.10043
2.45	3.55	-0.27427	-0.07347	-2.87986	0.08062	-0.20301
2.51	2.89	-0.19532	-0.29956	-2.14859	0.12789	-0.26302
4.70	3.48	1.36219	-0.09745	-0.07142	1.86505	1.16246
3.31	2.60	0.40076	-0.39891	-0.78309	0.31973	-0.36458
2.73	2.99	-0.04104	-0.26531	-1.72400	0.07207	-0.18956
3.34	2.81	0.41126	-0.32697	-0.67171	0.27604	-0.35532
6.07	2.04	2.43359	-0.59074	-0.23814	6.27133	11.51409

TABLE 2.1--Cont'd

HC	NOX	Z_1	Z_2	θ	r^2	$r^2 \log_e r^2$
3.03	2.62	0.19650	-0.39206	-1.10618	0.19232	-0.31706
4.64	3.54	1.31543	-0.07690	-0.05839	1.73627	0.95796
3.53	3.53	0.51050	-0.08032	-0.15606	0.26706	-0.35259
3.25	2.07	0.38566	-0.58047	-0.98437	0.48567	-0.35076
3.33	2.11	0.44156	-0.56677	-0.90894	0.51620	-0.34134
3.30	1.51	0.45199	-0.77230	-1.04130	0.80075	-0.17794
3.77	5.21	0.59451	0.49519	0.69450	0.59865	-0.30715
3.09	2.49	0.24702	-0.43659	-1.05591	0.25163	-0.34720
2.39	1.62	-0.21426	-0.73462	-1.85458	0.58558	-0.31338

TABLE 2.2 SORTED VALUES OF $U = 0.5 + \frac{\theta_k}{2\pi}$, THE ASSOCIATED CUMULATIVE FREQUENCY $F = \frac{k}{n+1}$, AND THE KS STATISTIC

θ	U	F	KS
-2.98999	0.02413	0.02083	0.00329
-2.97433	0.02662	0.04167	0.01505
-2.87986	0.04166	0.06250	0.02084
-2.71709	0.06756	0.08333	0.01577
-2.71189	0.06839	0.10417	0.03578
-2.62917	0.08155	0.12500	0.04345
-2.42186	0.11455	0.14583	0.03128
-2.37426	0.12212	0.16667	0.04454
-2.33517	0.12835	0.18750	0.05915
-2.15658	0.15677	0.20833	0.05156
-2.14859	0.15804	0.22917	0.07113
-1.98342	0.18433	0.25000	0.06567
-1.95123	0.18945	0.27083	0.08138
-1.85458	0.20483	0.29167	0.08683
-1.72400	0.22562	0.31250	0.08688
-1.71475	0.22709	0.33333	0.10624
-1.59263	0.24653	0.35417	0.10764
-1.55572	0.25240	0.37500	0.12260
-1.11736	0.32217	0.39583	0.07367
-1.10618	0.32395	0.41667	0.09272
-1.05591	0.33195	0.43750	0.10555
-1.04130	0.33427	0.45833	0.12406
-1.02553	0.33678	0.47917	0.14238
-0.98437	0.34333	0.50000	0.15667
-0.96627	0.34621	0.52083	0.17462
-0.91387	0.35455	0.54167	0.18711
-0.91111	0.35499	0.56250	0.20751
-0.90894	0.35534	0.58333	0.22800
-0.79500	0.37347	0.60417	0.23069
-0.78309	0.37537	0.62500	0.24963
-0.67171	0.39309	0.64583	0.25274
-0.60784	0.40326	0.66667	0.26341
-0.36885	0.44130	0.68750	0.24620
-0.24405	0.46116	0.70833	0.24718
-0.23814	0.46210	0.72917	0.26707
-0.15606	0.47516	0.75000	0.27484
-0.09777	0.48444	0.77083	0.28639

TABLE 2.2--Cont'd

θ	U	F	KS
-0.07142	0.48863	0.79167	0.30303
-0.05839	0.49071	0.81250	0.32179
0.04340	0.50691	0.83333	0.32643
0.04520	0.50719	0.85417	0.34697
0.16576	0.52638	0.87500	0.34862
0.43929	0.56992	0.89583	0.32592
0.69450	0.61053	0.91667	0.30613
1.04944	0.66702	0.93750	0.27048
1.14791	0.68270	0.95833	0.27564
1.85312	0.79493	0.97917	0.18423

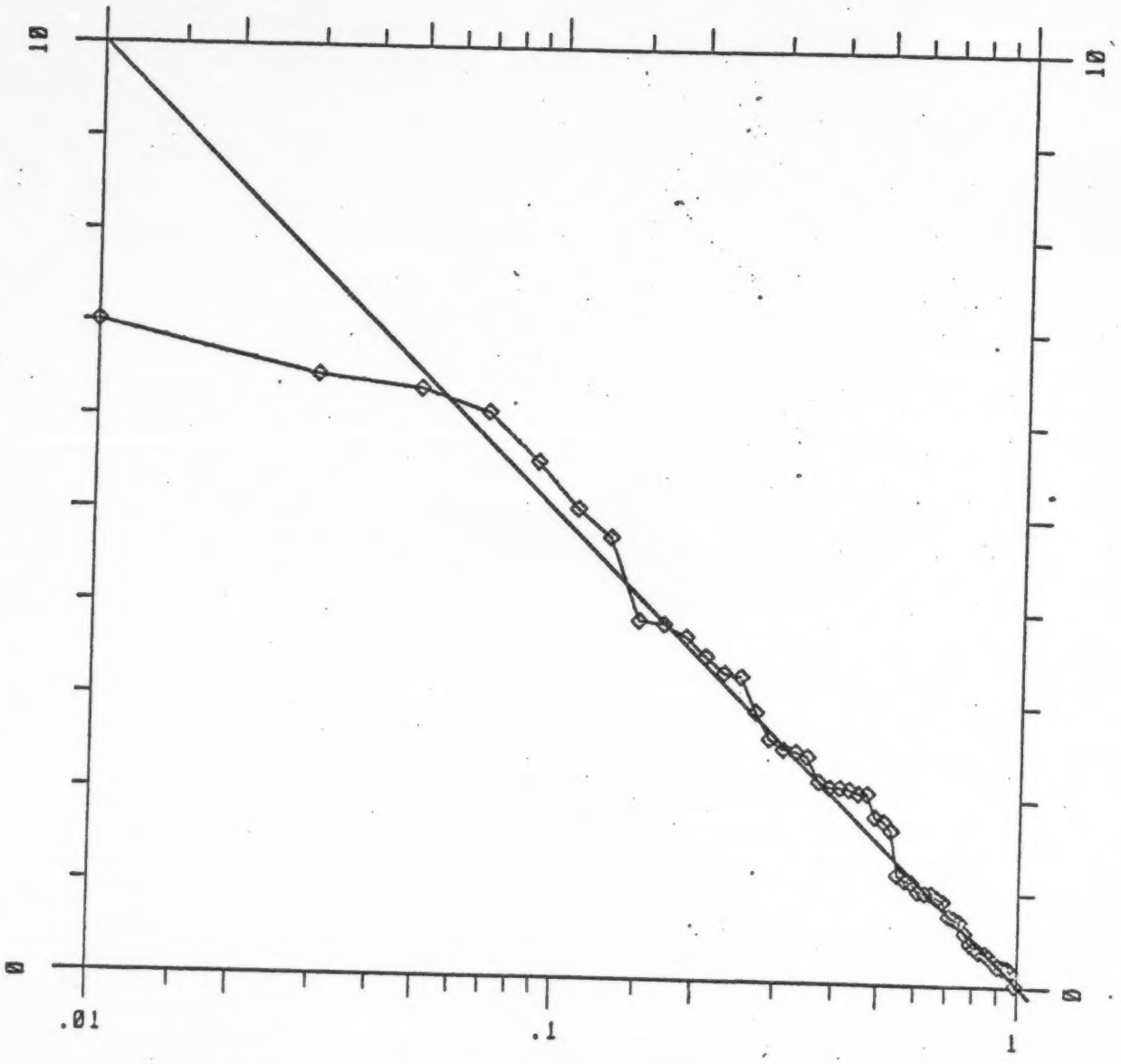


Figure 2.9a Distance representation of bivariate normal data on a semilog probability plot.

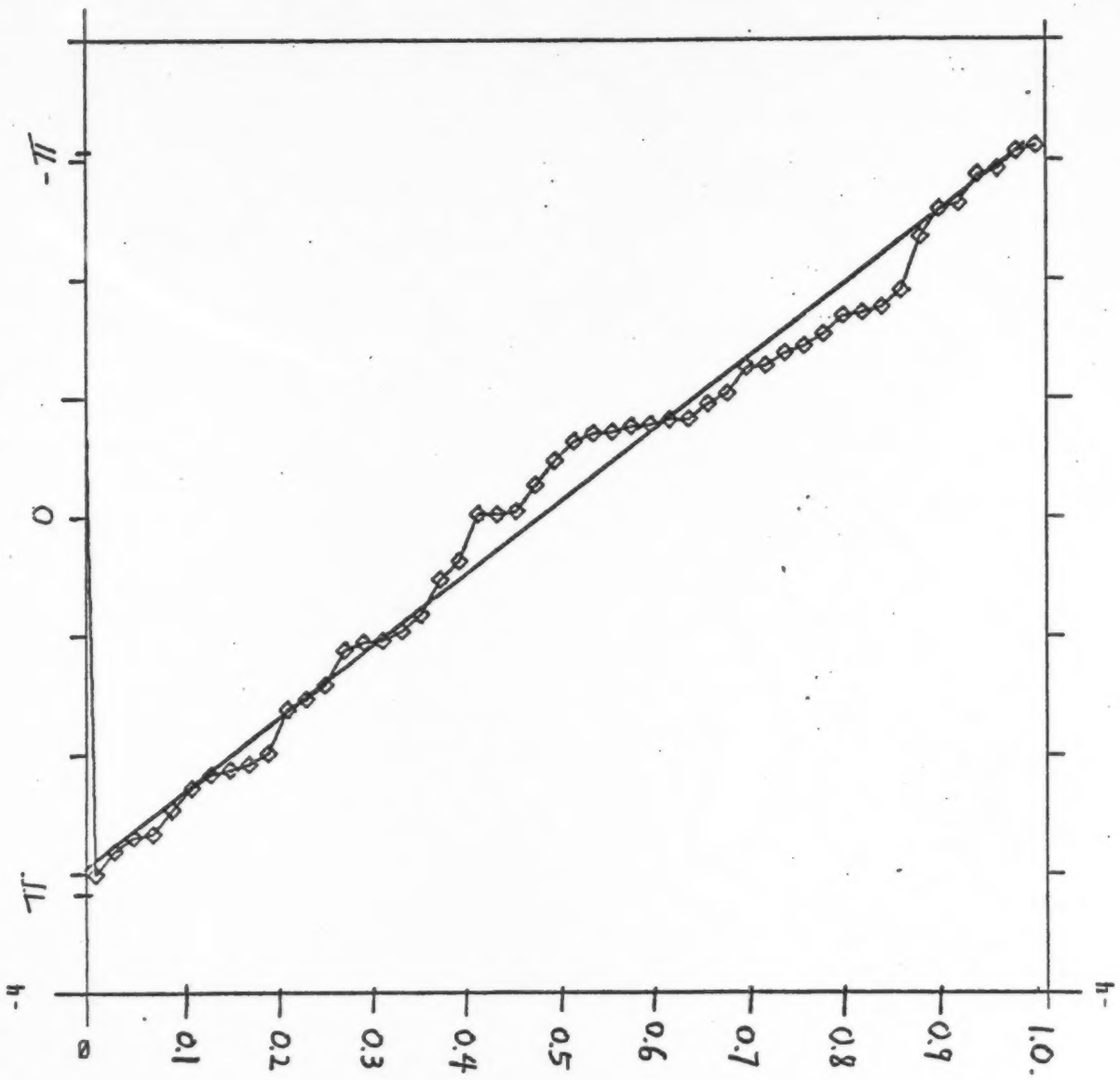


Figure 2.9b Angle representation of bivariate normal data on a uniform probability plot.

APPENDIX C

TESTS FOR NORMALITY

The following artical contains a test for normality that can be sued to test the hypothesis that each of the factors of the EFE is normally distributed.

© Journal of the American Statistical Association
March 1972, Volume 67, Number 337
Theory and Methods Section

An Approximate Analysis of Variance Test for Normality

S. S. SHAPIRO and R. S. FRANCIA*

This article presents a modification of the Shapiro-Wilk W statistic for testing normality which can be used with large samples. Shapiro and Wilk gave coefficients and percentage points for sample sizes up to 50. These coefficients required obtaining an approximation to the covariance matrix of the normal order statistics. The proposed test uses coefficients which depend only on the expected values of the normal order statistics which are generally available. Results of an empirical sampling study to compare the sensitivity of the test statistic to the W test statistic are briefly discussed.

1. THE W TEST FOR NORMALITY

The analysis of variance test statistic W developed in [5] for testing complete samples for normality is derived as follows:

Let $m' = (m_1, m_2, \cdots, m_n)$ denote the vector of expected values of standard normal order statistics, and let $V = (v_{ij})$ be the corresponding $n \times n$ covariance matrix. That is, if $x_1 \leq x_2 \leq \cdots \leq x_n$ denotes an ordered random sample of size n from a standard normal distribution ($\mu = 0$, $\sigma = 1$), then

$$E(x_i) = m_i (i = 1, 2, \cdots, n), \tag{1.1}$$

and

$$\mathrm{Cov}(x_i, x_j) = v_{ij} (i, j = 1, 2, \cdots, n). \tag{1.2}$$

Let $y' = (y_1, y_2, \cdots, y_n)$ denote a vector of ordered random observations. It is desired to derive a test for the composite hypothesis that $\{y_i\}$ is a sample from a normal distribution with unknown mean μ and unknown variance σ^2. If $\{y_i\}$ is a normal sample, then y_i may be expressed as

$$y_i = \mu + \sigma x_i (i = 1, 2, \cdots, n). \tag{1.3}$$

The W test statistic is then defined as

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{1.4}$$

where

$$a' = (a_1, a_2, \cdots, a_n) = \frac{m'V^{-1}}{(m'V^{-1}V^{-1}m)^{\frac{1}{2}}}. \tag{1.5}$$

To date, however, the elements of V are given only for sample sizes up to 20 in [4]. Approximations were developed in [5] for calculating coefficients for using the W test up to sample size 50.

2. AN APPROXIMATE W' TEST FOR NORMALITY

For large samples, the observations $\{y_i\}$ may be treated as if they were independent and as suggested in [1], the identity matrix I can be substituted for V^{-1} in the estimation of the slope of the regression line. Consequently, an approximate W' test for normality may be defined as

$$W' = \frac{(\sum_{i=1}^{n} b_i y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.1}$$

where

$$b' = (b_1, b_2, \cdots, b_n) = \frac{m'}{(m'm)^{\frac{1}{2}}}. \tag{2.2}$$

Values of m are given in [2] for sample sizes $n = 2(1)100(25)300(50)400$.

The null distribution of W' was approximated by an empirical sampling study. One thousand normal random sample were generated for each sample size $n = 35$, 50, 51(2)99, with the use of IBM subroutine GAUSS [3]. For each sample size, the 1,000 values of W' were calculated by equation (2.1), and the quantities $W'_j(p)$ were determined from the corresponding ordered W'_j using the relationship $p_j = j/n - 1$,

$$j = 10,50(50)200(300)800(50)950,990.$$

The table gives the empirical percentage points of W' for the above values of n and j.

3. COMPARISON OF THE SENSITIVITIES OF THE W AND W' TEST STATISTICS

An empirical sampling study was conducted to compare the sensitivities of the W and W' test statistics in detecting non-normality. One hundred samples of size $n = 35$ and 50 were generated randomly from 37 alternative distributions. The alternative distributions came from 12 different families and represented a wide variety of shapes including skewed and symmetric as well as

* S. S. Shapiro is director, Scientific Services, and R. S. Francia is statistician, with Programming Methods, Inc., 1301 Sixth Avenue, New York, N. Y.
The authors are indebted to Mrs. Ruth Kurtz and Mr. Tom Brady who programmed the routines for generating the statistical tables for the power study and the W' statistic, and to the editor and referees for their helpful comments.

EMPIRICAL PERCENTAGE POINTS OF THE APPROXIMATE W' TEST

n	α										
	.01	.05	.10	.15	.20	.50	.80	.85	.90	.95	.99
35	0.919	0.943	0.952	0.956	0.964	0.976	0.982	0.985	0.987	0.989	0.992
36	.935	.953	.963	.968	.971	.981	.987	.988	.990	.991	.994
37	0.935	0.954	0.964	0.968	0.971	0.981	0.988	0.989	0.990	0.992	0.994
38	.938	.957	.966	.969	.972	.982	.988	.989	.990	.992	.994
39	.940	.958	.965	.971	.973	.983	.988	.990	.991	.992	.994
40	.944	.961	.966	.971	.974	.983	.989	.991	.991	.992	.994
41	.945	.962	.967	.972	.975	.983	.989	.990	.991	.992	.994
42	0.947	0.963	0.968	0.973	0.975	0.984	0.990	0.990	0.991	0.992	0.994
43	.947	.966	.970	.973	.976	.984	.990	.991	.992	.993	.994
44	.948	.965	.971	.974	.976	.985	.990	.991	.992	.993	.995
45	.950	.966	.971	.974	.977	.985	.990	.991	.992	.993	.995
46	.951	.966	.972	.976	.978	.986	.990	.991	.992	.993	.995
47	0.953	0.967	0.972	0.975	0.978	0.986	0.990	0.991	0.992	0.994	0.995
48	.954	.968	.973	.976	.979	.986	.991	.992	.993	.994	.995
49	.954	.969	.973	.976	.979	.986	.991	.992	.993	.994	.995
50	.957	.969	.974	.977	.980	.987	.991	.992	.993	.994	.996
51	.957	.970	.975	.978	.980	.987	.991	.992	.993	.994	.996
52	0.958	0.970	0.975	0.979	0.981	0.987	0.992	0.992	0.993	0.994	0.996
53	.960	.971	.976	.979	.981	.988	.992	.992	.993	.994	.996
54	.961	.972	.977	.980	.981	.988	.992	.993	.994	.994	.996
55	.961	.973	.977	.980	.982	.988	.992	.993	.994	.994	.996
56	.961	.972	.977	.981	.982	.988	.992	.993	.994	.995	.996
57	0.962	0.973	0.978	0.981	0.983	0.989	0.992	0.993	0.994	0.995	0.996
58	.963	.973	.979	.981	.983	.989	.992	.993	.994	.995	.996
59	.965	.974	.979	.981	.983	.989	.993	.993	.994	.995	.996
60	.965	.973	.979	.982	.984	.989	.993	.993	.994	.995	.996
61	.967	.976	.980	.982	.984	.989	.993	.994	.994	.995	.996

continuous and discrete distributions. The W and the W' tests were performed on each sample. Using a 10 percent confidence level, a count was made of the number of times, out of a possible 100, that the W and W' tests rejected the hypothesis that the sample came from a normal population.

In general, the W' test appeared to be more sensitive than the W test when the alternative distribution was continuous and symmetric with high fourth moment (as compared to the normal distribution), when it was near normal and when it was discrete and skewed. The two tests appeared to be equivalent for alternative distributions which were continuous and asymmetric with high fourth moment and discrete and symmetric. The W test is superior to the W' test for other alternative distributions. Overall, the differentials in the power were small.

An earlier article [6] reported that for many alternative distributions, the W test has power as good or better than the following test procedures: $\sqrt{b_1}$, b_2, Kolmogorov-Smirnov, Cramer-Von Mises, Weighted Cramer-Von Mises, Durbin, chi-squared and David's u. Thus the W' test would also have this property.

4. CONCLUDING REMARKS

An approximate and simplified version of the W test was developed by estimating the slope of a regression line by simple least squares instead of the more precise procedure using generalized least squares. The operational constants for the W' test are easier to obtain and it is not limited to sample sizes less than 51. The sampling studies undertaken seem to indicate that the sensitivities of the W test and the approximate W' test were similar and any differences depended on the characteristics of the alternative distributions. Consequently, one was either less powerful, equal or more powerful than the other, depending on the nature of the alternative distribution. The use of the W' procedure for sample sizes greater than 50 is recommended and the extension to sample sizes over 100 can readily be accomplished. However, one must be careful in such extrapolations. In [7] and [8] it was shown that the chi-squared goodness-of-fit test had zero asymptotic efficiency relative to the Kolmogorov-Smirnov test while the results of [6] indicated that for sample size 50 for many alternative distributions the chi-squared procedure was the more powerful. This clearly cannot hold true as the sample size increases and a crossover point will eventually be reached.

[*Received June 1970. Revised June 1971*]

REFERENCES

[1] Gupta, A. K., "Estimation of the Mean and Standard Deviation of a Normal Population from a Censored Sample," *Biometrika*, 39 (December 1952), 260–73.

[2] Harter, H. L., "Expected Values of Normal Order Statistics," *Biometrika*, 48, Nos. 1 and 2 (June 1961), 151–65.

[3] IBM Application Program, Document No. H20-0205-3, IBM Corporation, Poughkeepsie, N.Y., 77.

[4] Sarhan, A.E., and Greenberg, B. G., "Estimation of Location and Scale Parameters by Order Statistics from Singly and Doubly Censored Samples Part I," *Annals of Mathematical Statistics*, 27 (June 1956), 427–51.

[5] Shapiro, S. S. and Wilk, M. B., "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52 (December 1965), 591–611.

[6] ———, Wilk, M. B., and Chen, H. J., "A Comparative Study of Various Tests for Normality," *Journal of the American Statistical Association*, 63 (December 1968), 1343–72.

[7] Wald, A., and Wolfowitz, J., "Confidence Limits for Continuous Distribution Functions," *Annals of Mathematical Statistics*, 10 (March 1939), 105–18.

[8] ——— and Wolfowitz, J., "Notes on Confidence Limits for Continuous Distribution Functions," *Annals of Mathematical Statistics*, 12 (March 1941), 118–19.